Structure From Motion

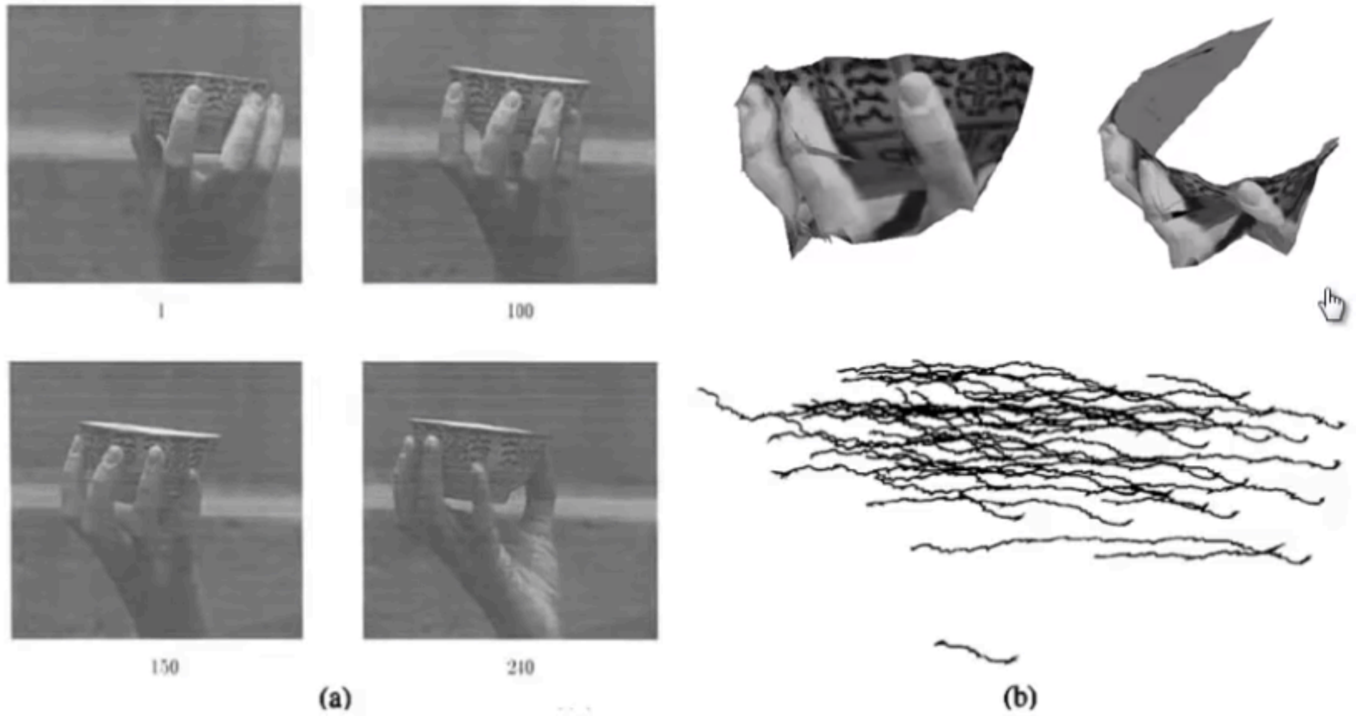
Computer Vision (CS0029)

PhotoSyth

- https://www.youtube.com/watch?v=p16frKJLVi0

Structure from Motion Problem

- Given point correspondences, compute 3-D motion (rotation, translation) and shape (depth)



Images KLT Tracks

Structure from Motion

- **Tomasi and Kanade Factorization**
  - https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=E114ACADDC394E7F794F6DC87F3F817B?doi=10.1.1.131.9807&rep=rep1&type=pdf

- Assumptions
  - The camera model is orthographic projection
  - The positions of “P” points in “F” frames (F>=3), which are not all coplanar, and have been tracked
  - The entire sequence has been acquired before starting (batch mode)
  - Camera calibration not needed, if we accept 3D points up to a scale factor

Feature Points

- Image points: $\{(u_{fp}, v_{fp}) | f = 1, \ldots, F, p = 1, .., P\}$

- $$W = \begin{bmatrix} u_{11} & \ldots & u_{1p} \\ . & . & . \\ . & . & . \\ u_{F1} & \ldots & u_{FP} \\ v_{11} & \ldots & v_{1P} \\ . & . & . \\ . & . & . \\ v_{F1} & \ldots & v_{FP} \end{bmatrix}$$

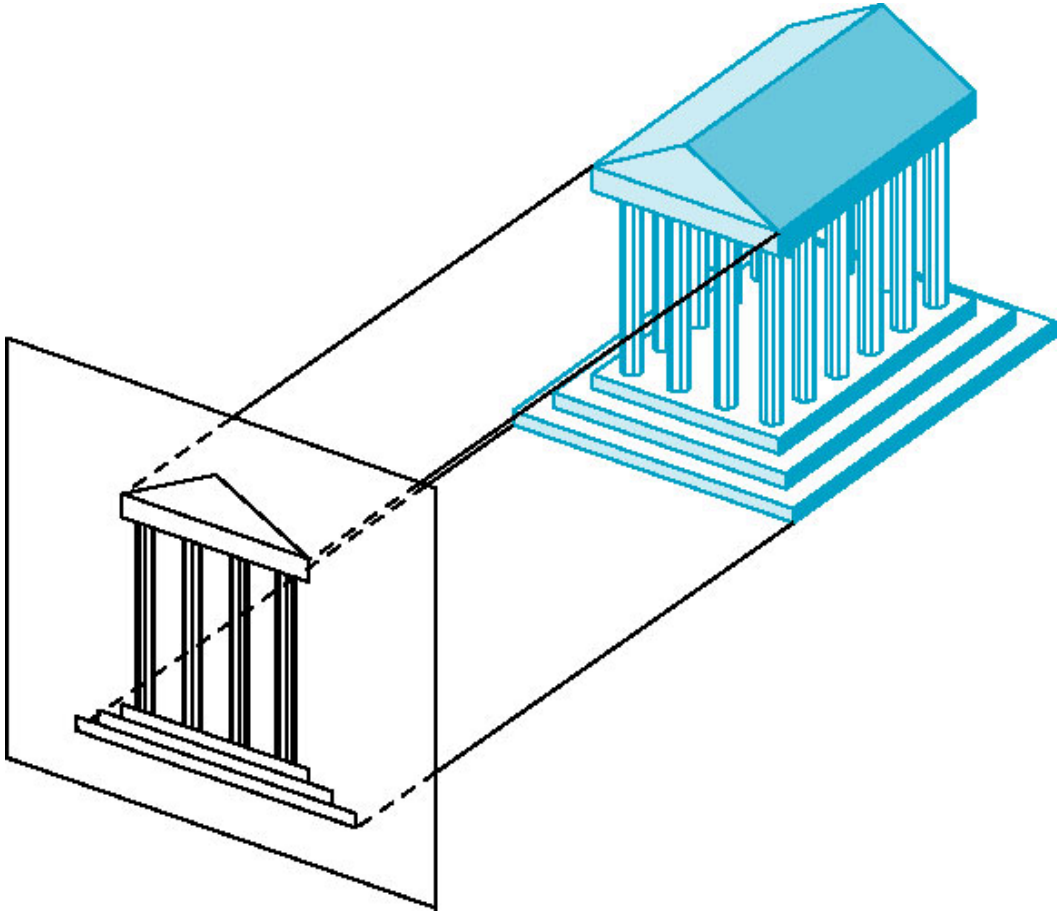
$$\mathrm{W} = \begin{bmatrix} U \\ - \\ V \end{bmatrix}$$

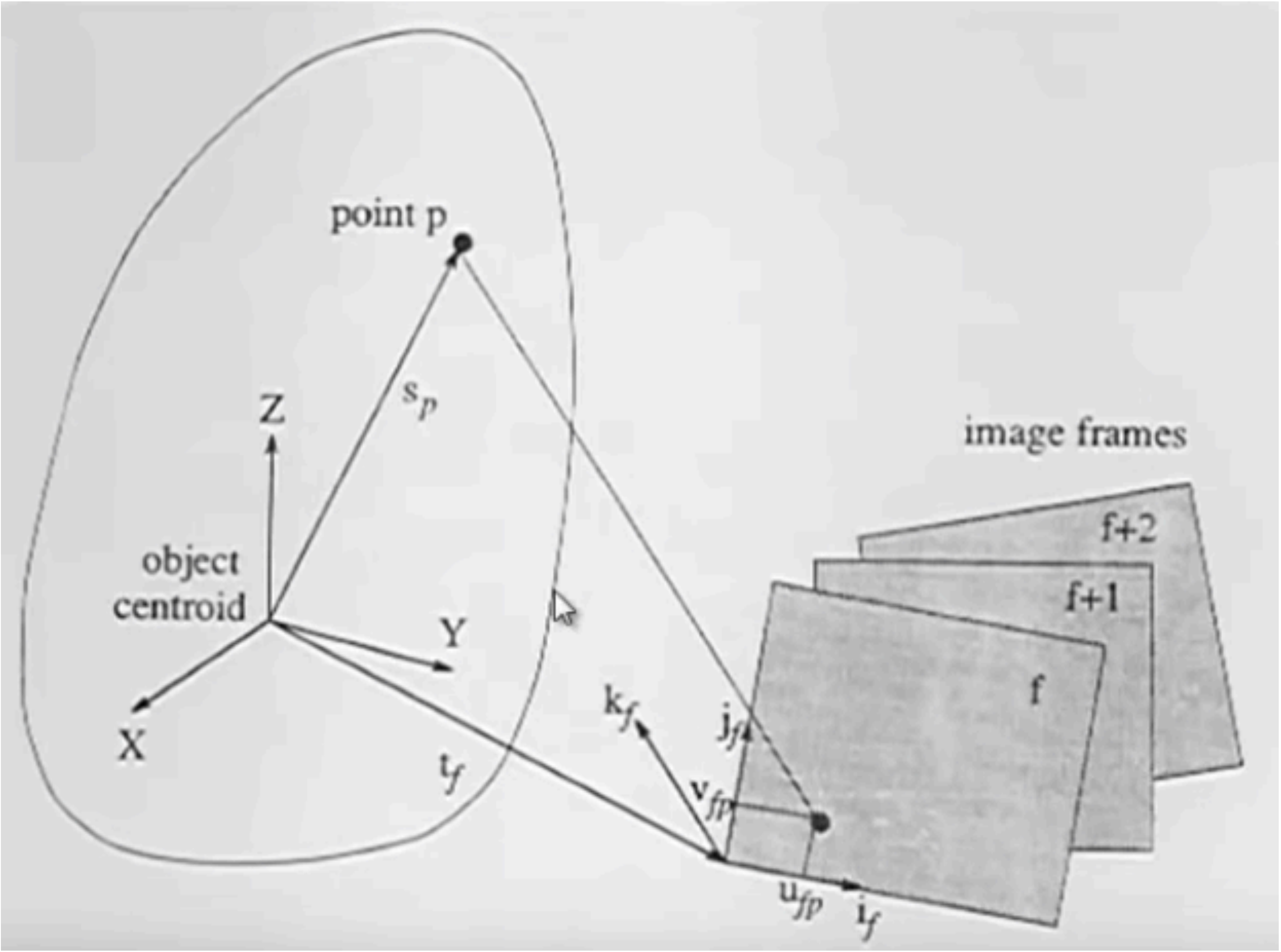
Mean Normalize Feature Points

- Average u, v of all points on image f
  - $a_f = \frac{1}{P}\sum_{p=1}^{P} u_p$
  - $b_f = \frac{1}{P}\sum_{p=1}^{P} v_p$

- Normalize
  - $\widetilde{u_{fP}} = u_{fP} - a_f$
  - $\widetilde{v_{fP}} = v_{fP} - b_f$

Orthographic Projection





Orthographic Projection

- $S_p = (X_p, Y_p, Z_p)$. : 3D world point of point p

- $k_f = i_f \times j_f$ $\quad\quad$ i, j, k are unit vector (of camera frame) along X, Y, Z

- Orthographic Projection
  - $u_{fP} = i_f^T(s_p - t_f)$
  - $v_{fP} = j_f^T(s_p - t_f)$

$$\widetilde{u_{fP}} = u_{fP} - a_{fP} = i_f^T(s_p - t_f) - \frac{1}{P}\sum_{q=1}^{P} u_q$$

$$= i_f^T(s_p - t_f) - \frac{1}{P}\sum_{q=1}^{P} i_f^T(s_q - t_f) = i_f^T\left[s_P - \frac{1}{P}\sum_{q=1}^{P} s_q\right]$$

- **If origin of world is at the centroid of object points, second term is zero**

- $\widetilde{u_{fP}} = i_f^T\ s_P$
- $\widetilde{v_{fP}} = j_f^T\ s_P$

- $\widetilde{W} = \begin{bmatrix} \widetilde{U} \\ - \\ \widetilde{V} \end{bmatrix}$

- $\widetilde{u_{fP}} = i_f^T\ s_P$
- $\widetilde{v_{fP}} = j_f^T\ s_P$

- $\widetilde{W} = \begin{bmatrix} i_1^T \\ . \\ . \\ i_F^T \\ j_1^T \\ . \\ . \\ j_F^T \end{bmatrix} \begin{bmatrix} s_1 & \cdots & s_P \end{bmatrix} = RS$

2F*3

3*P

Rank of S is 3, because points in 3D space are not co-plannar

- $\widetilde{W} = \begin{bmatrix} \widetilde{U} \\ - \\ \widetilde{V} \end{bmatrix}$ $\quad \widetilde{W} = \begin{bmatrix} \widetilde{u_{11}} & \cdots & \widetilde{u_{1p}} \\ . & . & . \\ . & . & . \\ \widetilde{u_{F1}} & \cdots & \widetilde{u_{Fp}} \\ \widetilde{v_{11}} & \cdots & \widetilde{v_{1p}} \\ . & . & . \\ . & . & . \\ \widetilde{v_{F1}} & \cdots & \widetilde{v_{Fp}} \end{bmatrix}$

Rank Theorem

- Without noise, the registered measurement matrix is at most of rank three.

- $\widetilde{W} = \begin{bmatrix} i_1^T \\ . \\ . \\ i_F^T \\ j_1^T \\ . \\ . \\ j_F^T \end{bmatrix} \begin{bmatrix} s_1 & \cdots & s_P \end{bmatrix} = RS$

- Because W is a product of two matrices. The maximum rank of S is 3.

How to find Translation?

- $\widetilde{u_{fp}} = u_{fp} - a_f$
- $u_{fp} = \widetilde{u_{fp}} + a_f$ $\qquad\qquad \widetilde{u_{fp}} = i_f^T s_p$

- $u_{fp} = i_f^T s_p + a_f$ $\qquad\qquad u_{fp} = i_f^T (s_p - t_f)$

- Compare above two equations
  - $a_f = -i_f^T t_f$ $\qquad$ (a_f is projection of camera translation along x-axis)

How to Find Translation?

- $u_{fp} = i_f s_p + a_f$ $\qquad\qquad v_{fp} = j_f s_p + b_f$
- $\underset{2F*P}{W} = \underset{2F*3}{R}\underset{3*P}{S} + \underset{2F*1}{t}\underset{2F*1}{e_p^T}$ $\qquad\qquad a_f = -i_f^T t_f$

- $t = \left(a_1, \ldots, a_f, b_1, \ldots, b_f\right)^T$
- $e_p^T = (1, \ldots, 1)$

How to Find Translation

- Projected camera translation can be computed:

  - $-i_f^T t_f = a_f = \frac{1}{P}\sum_{p=1}^{P} u_p$

  - $-j_f^T t_f = b_f = \frac{1}{P}\sum_{p=1}^{P} v_p$

- $W = RS + te_p^T$

Noisy Measurements

- Without noise, the matrix $\widetilde{W}$ must be at most of rank 3. When noise corrupts the images, however, $\widetilde{W}$ will not be of rank 3. Rank theorem can be extended to the case of noisy measurements.

Singular Valued Decomposition (SVD)

- SVD $\underset{\text{2F*P}}{\widetilde{W}} = \underset{\text{2F*P}}{O_1}\,\underset{\text{P*P}}{\Sigma}\,\underset{\text{P*P}}{O_2}$

- Theorem: Any m by n matrix A, for which $m \geq n$, can be written as
  - $\underset{\text{m*n}}{A} = \underset{\text{m*n}}{O_1}\,\underset{\text{n*n}}{\Sigma}\,\underset{\text{n*n}}{O_2}$
  - Σ: is diagonal
  - O_1, O_2 are orthogonal. $O_1^T O_1 = O_2^T O_2 = I$

Approximate Rank

- $\widetilde{W} = O_1 \Sigma O_2$

- $O_1 = \begin{array}{cc} 3 & P\text{-}3 \\ \multicolumn{2}{c}{[O_1' O_1'']} \end{array} \quad 2F$

- $\Sigma = \begin{array}{c} \begin{array}{cc} 3 & P\text{-}3 \end{array} \\ \begin{bmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{bmatrix} \end{array} \quad 2F$ (without noise: Σ'' should be 0)

- $O_2 = \begin{array}{c} P \\ \begin{bmatrix} O_2' \\ O_2'' \end{bmatrix} \end{array} \quad \begin{array}{c} 3 \\ P\text{-}3 \end{array}$

$$O_1 \Sigma O_2 = O_1' \Sigma' O_2' + O_1'' \Sigma'' O_2''$$

Approximate Rank

- $\widetilde{W} = O_1'\Sigma'O_2' + \ O_1''\Sigma''O_2''$
- $\widetilde{W} = O_1'\Sigma'O_2'$
- The best rank 3 approximation to the ideal registered measurement matrix

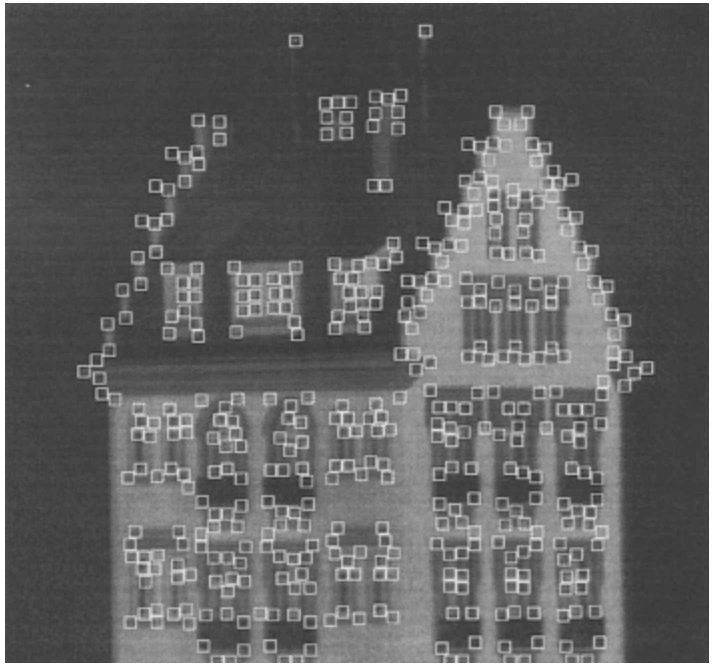
- Rank theorem for noisy measurement
  - The best possible shape and rotation estimate is obtained by considering only 3 greatest singular values of $\widetilde{W}$ together with the corresponding left, right eigenvectors.

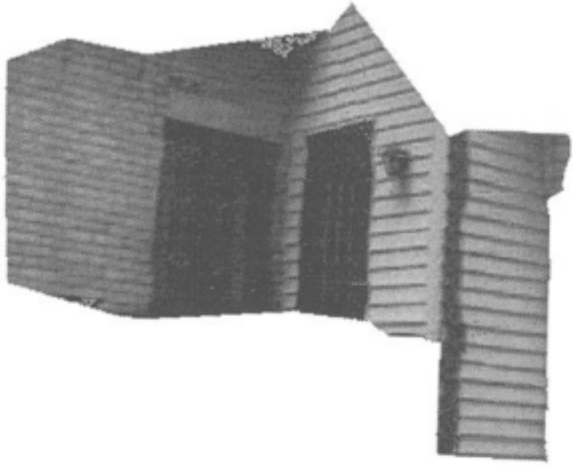
Approximate Rank

- $\widetilde{\mathrm{R}} = O_1'[\Sigma']^{\frac{1}{2}}$ Approximate Rotation matrix
- $\tilde{\mathrm{S}} = [\Sigma']^{\frac{1}{2}} O_2'$ Approximate Shape matrix
- $\widetilde{\mathrm{W}} = \widetilde{\mathrm{R}}\,\tilde{\mathrm{S}}$ This decomposition is not unique

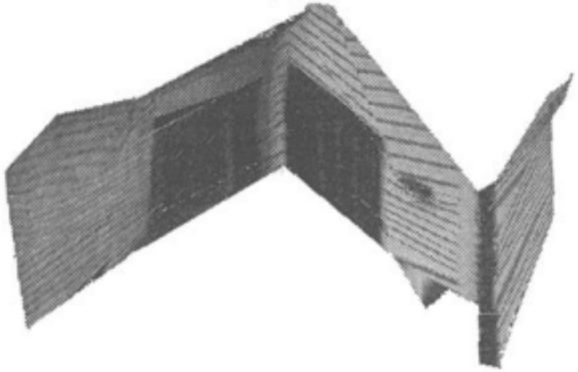
- $\widetilde{\mathrm{W}} = (\widetilde{\mathrm{R}}Q)(Q^{-1}\tilde{\mathrm{S}})$ Q is any 3*3 invertible matrix

Result





(a)



(b)

Summary

- More advance and better SfM approaches now. This is a basic one.

- Orthographic projection model is used

- $$\begin{bmatrix} \widetilde{u_{11}} & \cdots & \widetilde{u_{1p}} \\ . & . & . \\ . & . & . \\ \widetilde{u_{F1}} & \cdots & \widetilde{u_{Fp}} \\ \widetilde{v_{11}} & \cdots & \widetilde{v_{1p}} \\ . & . & . \\ . & . & . \\ \widetilde{v_{F1}} & \cdots & \widetilde{v_{Fp}} \end{bmatrix} = W = \overset{R}{\begin{bmatrix} i_1^T \\ . \\ . \\ i_F^T \\ j_1^T \\ . \\ . \\ j_F^T \end{bmatrix}} \overset{S}{[s_1 \quad \cdots \quad s_P]} \overset{SVD}{=} [O_1' O_1''] \begin{bmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{bmatrix} \begin{bmatrix} O_2' \\ O_2'' \end{bmatrix} \cong O_1' \Sigma' O_2' = O_1' [\Sigma']^{\frac{1}{2}} [\Sigma']^{\frac{1}{2}} O_2' = \tilde{\mathrm{R}}\, \tilde{\mathrm{S}}$$

$keep\ 3\ maximum$
$sigular\ values$