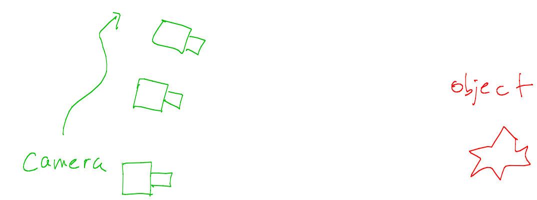
lecture 21

structure from motion

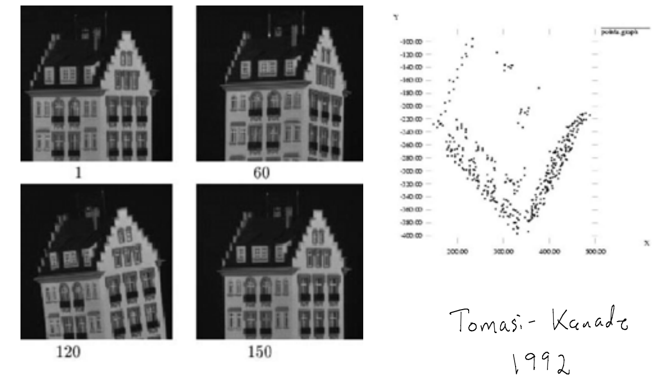
factorization method

(Tomasi - Kanade 1992)

"SFM": given F image frames with N points, can we estimate the 3D positions of points ("structure") and the positions of camera ("motion").



Today we look at a specific version of this (general) problem.



Tomasi - Kanade 1992

SFM Factorization Methods

Tomasi - Kanade (1992)

- orthographic projection
- rigid object
- all points are visible in each frame

→ multiple moving rigid objects

→ non-rigid objects (faces)

→ perspective projection models

→ missing points and/or outliers

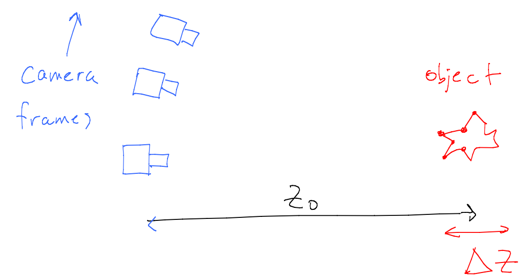
- F images ("frames" of video)
- N corresponding points

$$\{ (x_{ki}, y_{ki}) : i \in 1, .. N, \ k = 1, .. F \}$$

track keypoints from frame to frame

Assume Z_0 is approximately the same for each frame, but unknown.

Recall weak perspective i.e. $\Delta Z \ll Z_0$.

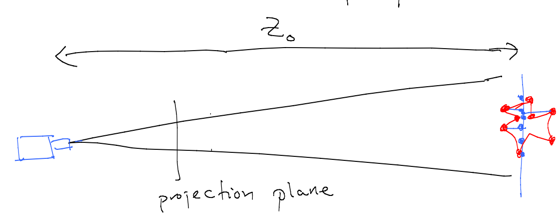


Tomasi - Kanade's 2 steps

- set up the problem: make assumptions about the geometry
- solve for the orientation of the cameras and the 3D structure

(This is the clever and more interesting part)

Recall weak perspective



$$\begin{bmatrix} wx \\ wy \\ w \end{bmatrix} = P \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & Z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Tomasi - Kanade use an simpler projection model - called "Orthographic Projection".

$$x \rightarrow, \quad y \rightarrow \begin{pmatrix} X \\ Y \\ 0 \end{pmatrix} \qquad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

Orthographic Projection

$$\begin{bmatrix} x_{ik} \\ y_{ik} \end{bmatrix} = \tilde{R}_k \, [I \mid -C_k] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}$$

2x3 matrix
The two rows are the X and Y axes of camera k, written in scene coordinate system

$$\begin{bmatrix} x_{ik} \\ y_{ik} \end{bmatrix} = \tilde{R}_k \, [I \mid -C_k] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}$$

$2 \times 1 \quad 2 \times 3 \quad 3 \times 4 \quad 4 \times 1$

Huh?

Pixel units same as scene units?
(Not a problem if we don't know Z_0)

Tomasi Kanade

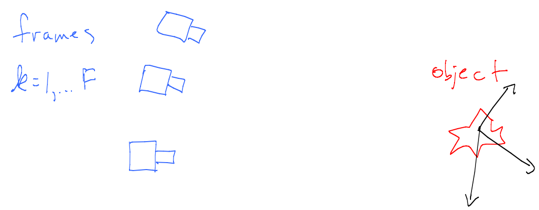
- set up the problem
  - orthographic projection
  - → shift $(x_{ik}\ y_{ik})$ in image k so mean position is (0,0)
- Solve the problem

$$\begin{bmatrix} x_{ik} \\ y_{ik} \end{bmatrix} = \tilde{R}_k \, [I \mid -C_k] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}$$

$$\sum_{i=1}^{N} \begin{bmatrix} x_{ik} \\ y_{ik} \end{bmatrix} = \tilde{R}_k \underbrace{\sum_{i=1}^{N} \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}}_{\text{Assume } = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}} - N \tilde{R}_k C_k$$

without loss of generality

$$\begin{bmatrix} x_{ik} - \bar{x}_k \\ y_{ik} - \bar{y}_k \end{bmatrix} = \tilde{R}_k \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}$$



Summary thus far:

We reduced the problem to

- set of pixel positions $(x_{ik}\ y_{ik})$ shifted so mean in frame k is (0,0)
- set of unknown $X_i Y_i Z_i$ scene points and unknown camera orientations (defined by first 2 rows of R_k)

TASK: Solve for $X_i Y_i Z_i$ $\tilde{R}_k$

Notation: for each of matrix elements below, we also subtract the mean ie. $\bar{x}_k$ or $\bar{y}_k$

N key points (x,y) per frame

$$A = \begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1N} \\ y_{11} & \cdots & y_{1i} & \cdots & y_{1N} \\ \vdots & & \vdots & & \vdots \\ x_{F1} & \cdots & x_{Fi} & \cdots & x_{FN} \\ y_{F1} & \cdots & y_{Fi} & \cdots & y_{FN} \end{bmatrix} \quad 2F$$

$$\begin{bmatrix} x_{ik} - \bar{x}_k \\ y_{ik} - \bar{y}_k \end{bmatrix} = \tilde{R}_k \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}$$

$$A = \begin{bmatrix} x_{ik} - \bar{x}_k \\ y_{ik} - \bar{y}_k \end{bmatrix}_{2F \times N}$$

$$A = \underbrace{\begin{bmatrix} \tilde{R}_k \end{bmatrix}}_{2F \times 3} \underbrace{\begin{bmatrix} X_1 & \cdots & X_N \\ Y_1 & \cdots & Y_N \\ Z_1 & \cdots & Z_N \end{bmatrix}}_{3 \times N}$$

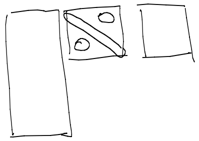
Tomasi-Kanade Factorization

$$A = \underbrace{\begin{bmatrix} \tilde{R}_k \end{bmatrix}}_{2F \times 3} \underbrace{\begin{bmatrix} X_1 & \cdots & X_N \\ Y_1 & \cdots & Y_N \\ Z_1 & \cdots & Z_N \end{bmatrix}}_{3 \times N} + \text{noise}$$

A is a rank 3 matrix + noise.

We want to factor it into "motion" $\tilde{R}$ and structure $\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$

$$A = U \Sigma V^T =$$



$2F \times N \quad 2F \times N \quad N \times N \quad N \times N$

Fact from linear algebra: the best least squares rank r approximation to a matrix A is obtained by setting $\sigma_{r+1}, \dots, \sigma_N$ to 0.

(In our case, $r = 3$.)

$$A = U \Sigma V^T$$

Let $\tilde{A}$ be a rank 3 matrix of same dimensions as A.

$$\| A - \tilde{A} \| \equiv \sum_{k,i} \left(A_{ki} - \tilde{A}_{ki} \right)^2$$

i.e. find the $\tilde{A}$ that minimizes the least squared error.

Called the Frobenius norm.

$$\| A - \tilde{A} \|$$

$$= \| U \Sigma V^T - U U^T \tilde{A} V V^T \|$$

not difficult to show

$$= \| \Sigma - U^T \tilde{A} V \|$$

This is minimized when $U^T \tilde{A} V$ is diagonal with $\sigma_r \dots \sigma_N = 0$.

Compute $A = U \Sigma V^T$

$2F \times N \quad N \times N \quad N \times N$

Set $\tilde{A} = \underset{2F \times 3}{\tilde{U}} \underset{3 \times 3}{\begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix}} \underset{3 \times N}{\tilde{V}^T}$

$$\stackrel{?}{=} \tilde{R} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

Problem: the rows of $\tilde{U}$ are unlikely to be orthogonal and of unit length.

Solution (trick):

$$\tilde{A} = \underbrace{\underset{2F \times 3}{\tilde{U}} \underset{3 \times 3}{Q}} \underset{3 \times 3}{Q^{-1}} \underset{3 \times 3}{\begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix}} \underset{3 \times N}{\tilde{V}^T}$$

Find invertible matrix Q such that

$$\tilde{R}_k \approx \tilde{U}_k Q,$$

i.e. N pairs of orthonormal row vectors.

Details

Notation $\tilde{U}_k = \begin{bmatrix} u_{2k-1} \\ u_{2k} \end{bmatrix}_{2 \times 3}$

Solve non-linear least squares:

$$(u_{2k-1} Q)^T u_{2k-1} Q = 1$$

$$(u_{2k} Q)^T (u_{2k} Q) = 1$$

$$(u_{2k-1} Q)^T (u_{2k} Q) = 0$$

3F non-linear constraints

SFM Factorization Methods

Tomasi-Kanade

- multiple moving rigid objects ("A" has multiple parts)
- non-rigid objects (faces) ("structure" is a time varying linear combination)
- perspective projection models (more non-linearities, "bundle adjustment")
- missing points and/or outliers