

Structure from motion

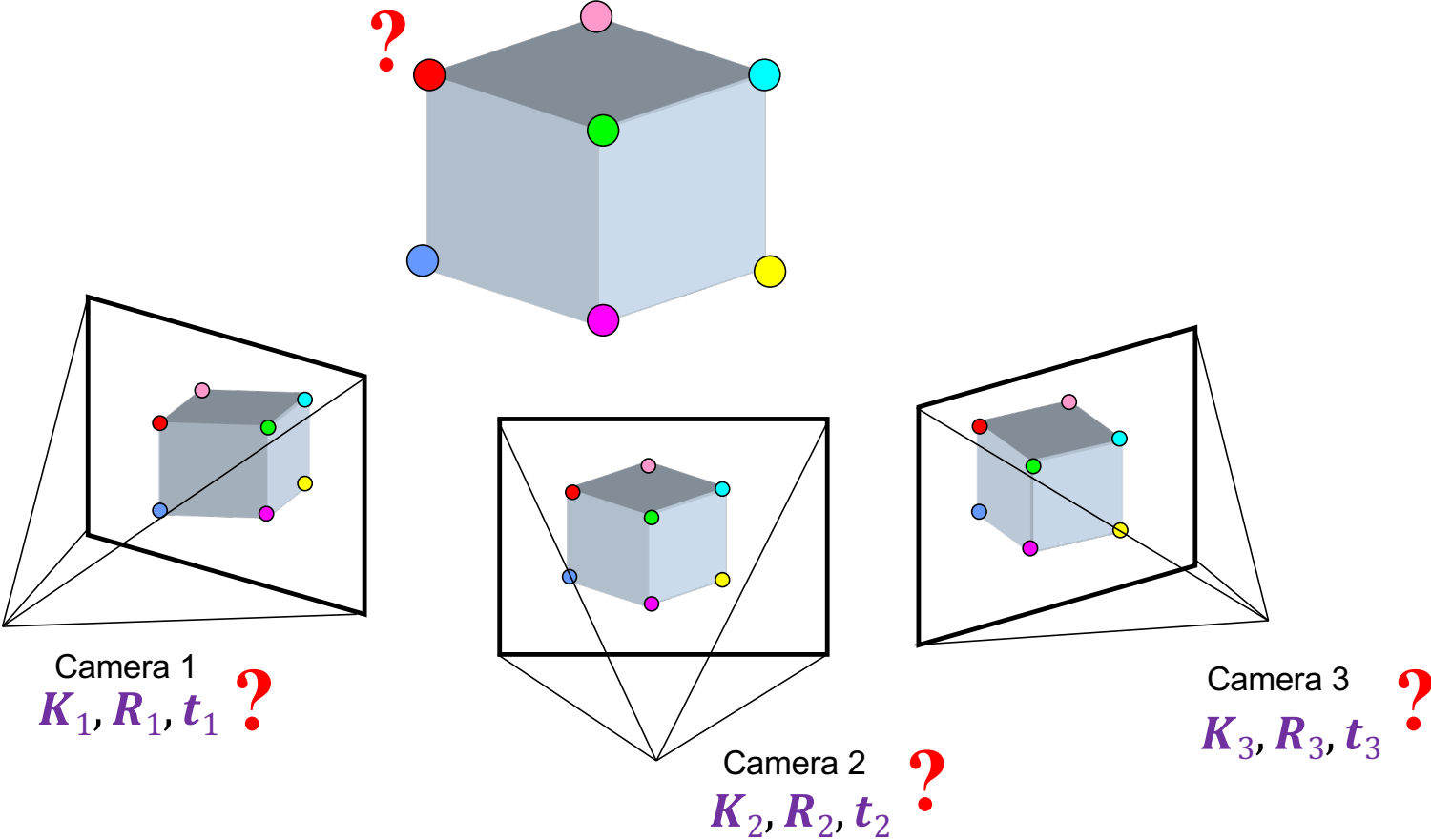


Драконъ, видимый подъ различными углами зрѣнія
По гравюру на мѣди наз. „Oculus artificialis teleiopicus“ Нана. 1702 года.

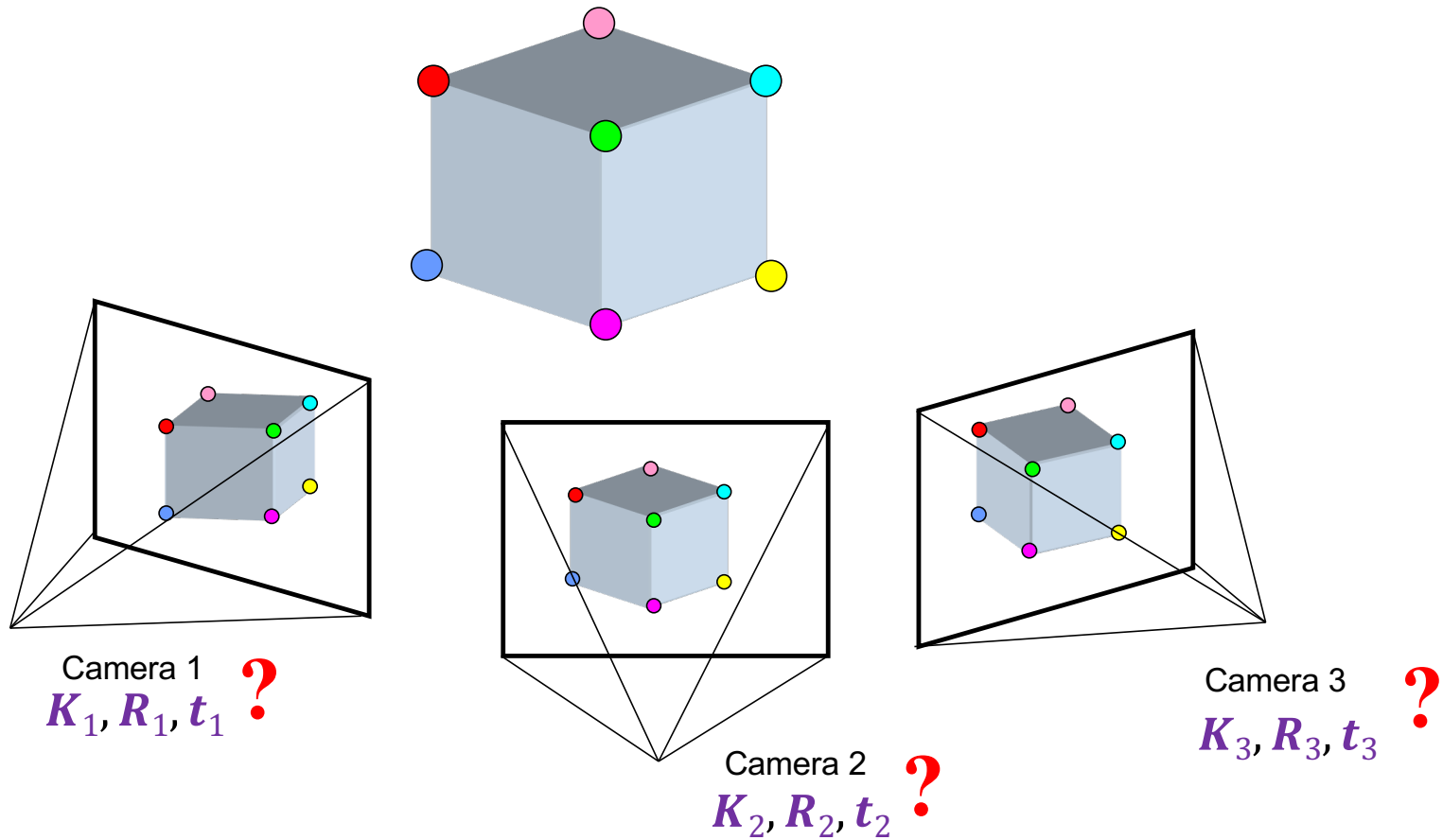
Outline: Structure from motion

- Problem definition and ambiguities
- Affine structure from motion
 - Factorization
- Projective structure from motion
 - Bundle adjustment
- Modern structure from motion pipeline

Structure from motion

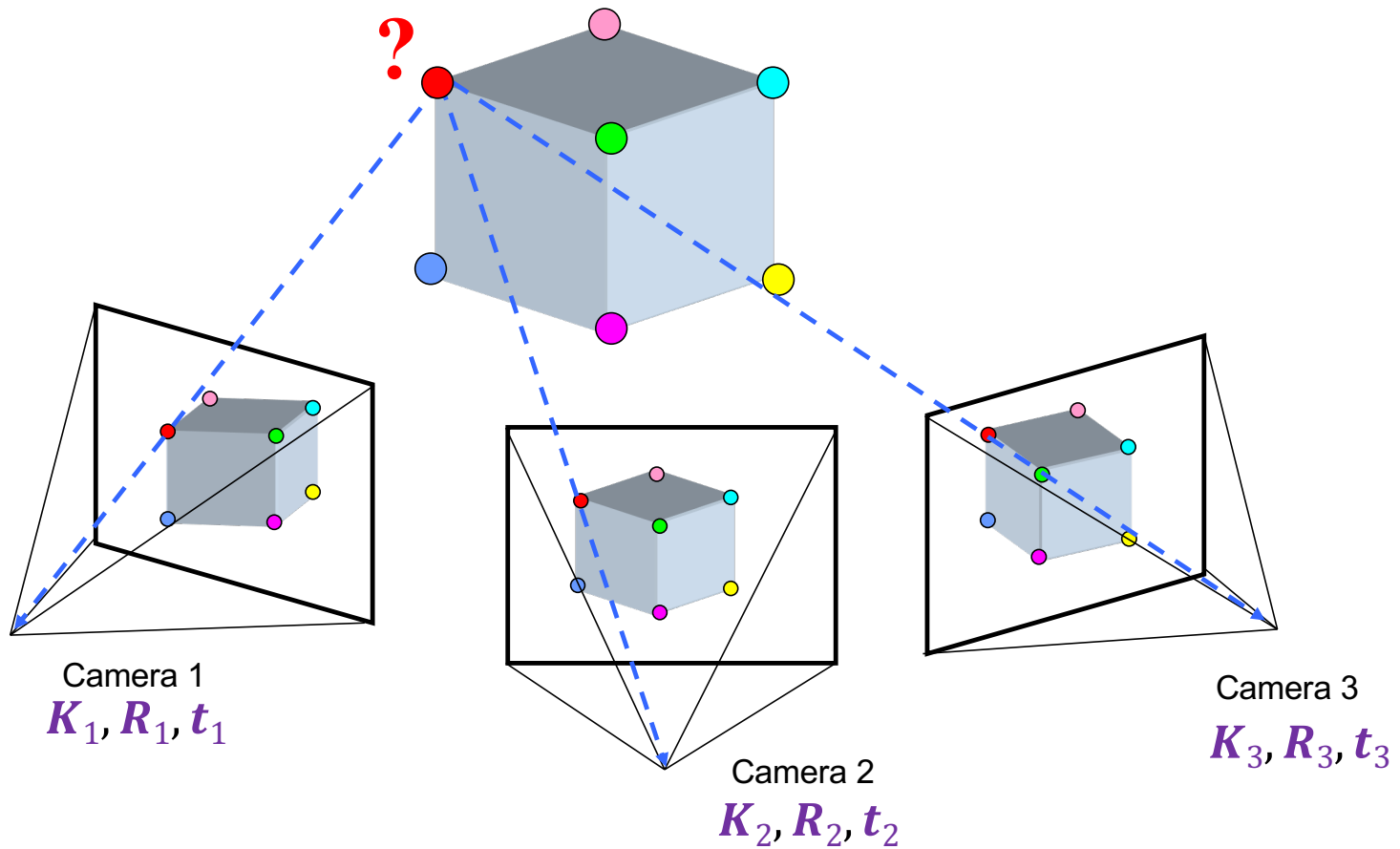


Recall: Calibration



- Given a set of *known* 3D points seen by a camera, compute the camera parameters

Recall: Triangulation



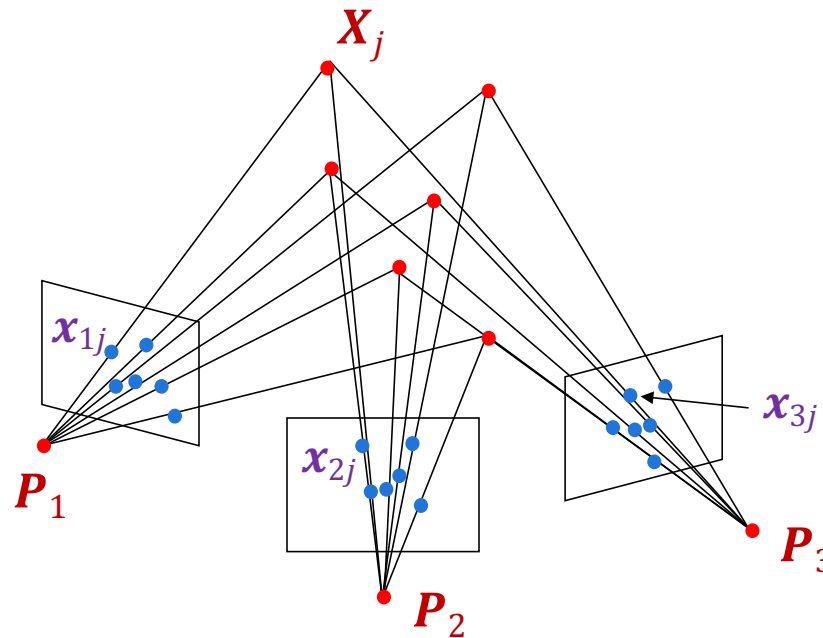
- Given *known cameras* and projections of the same 3D point in two or more images, compute the 3D coordinates of that point

Structure from motion: Problem formulation

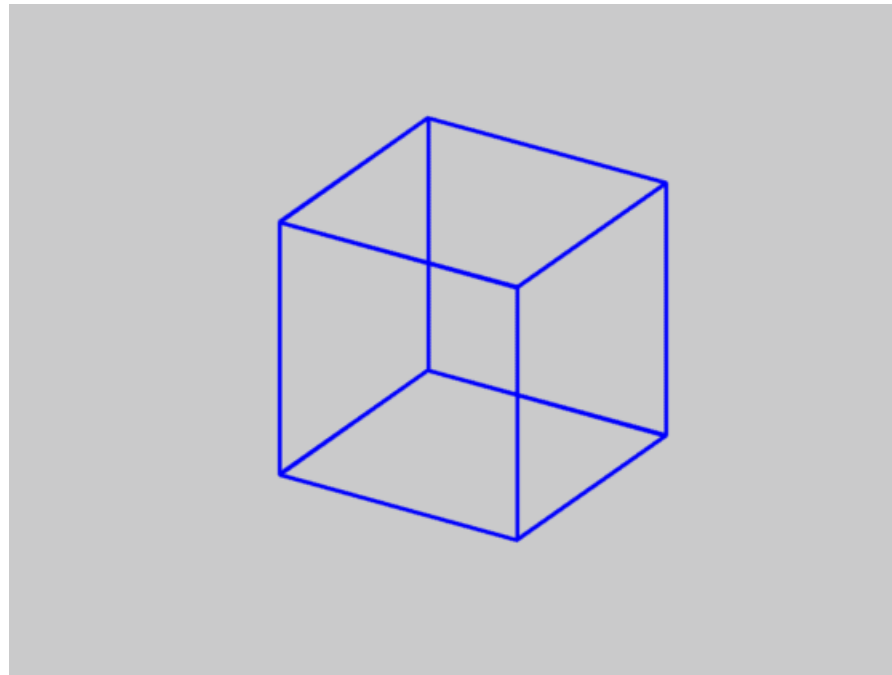
- Given: m images of n fixed 3D points such that (ignoring visibility)

$$\mathbf{x}_{ij} \cong \mathbf{P}_i \mathbf{X}_j, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

- Problem: estimate m projection matrices \mathbf{P}_i and n 3D points \mathbf{X}_j from the mn correspondences \mathbf{x}_{ij}



Is SFM always uniquely solvable?

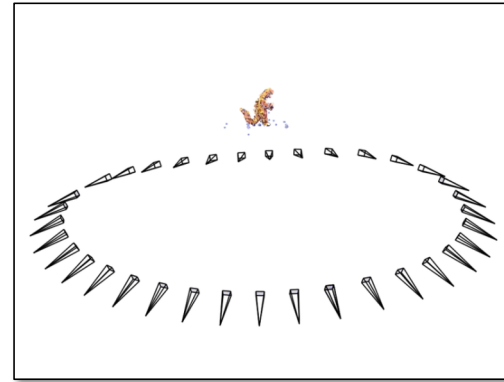
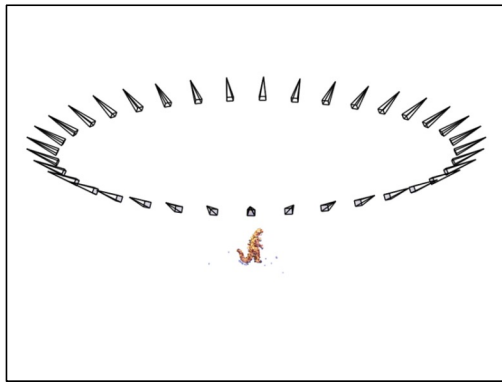


Necker cube

Source: N. Snavely

Is SFM always uniquely solvable?

- Could actually happen in affine structure from motion:



Source: N. Snavely

Structure from motion ambiguity

- If we scale the entire scene by some factor k and, at the same time, scale the camera matrices by the factor of $1/k$, the projections of the scene points remain exactly the same:

$$x \cong PX = \begin{pmatrix} 1 \\ -P \\ k \end{pmatrix} (kX)$$

- Without a reference measurement, it is impossible to recover the absolute scale of the scene!
- In general, if we transform the scene using a transformation Q and apply the inverse transformation to the camera matrices, then the image observations do not change:

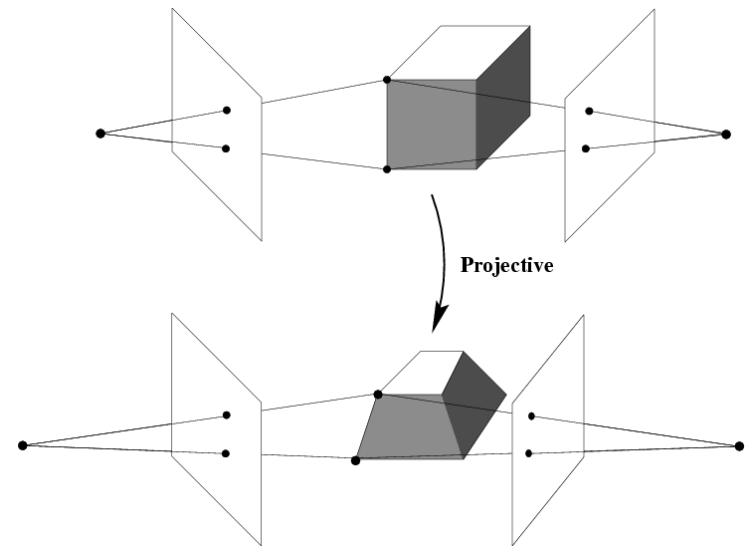
$$x \cong PX = (PQ^{-1})(QX)$$

Projective ambiguity

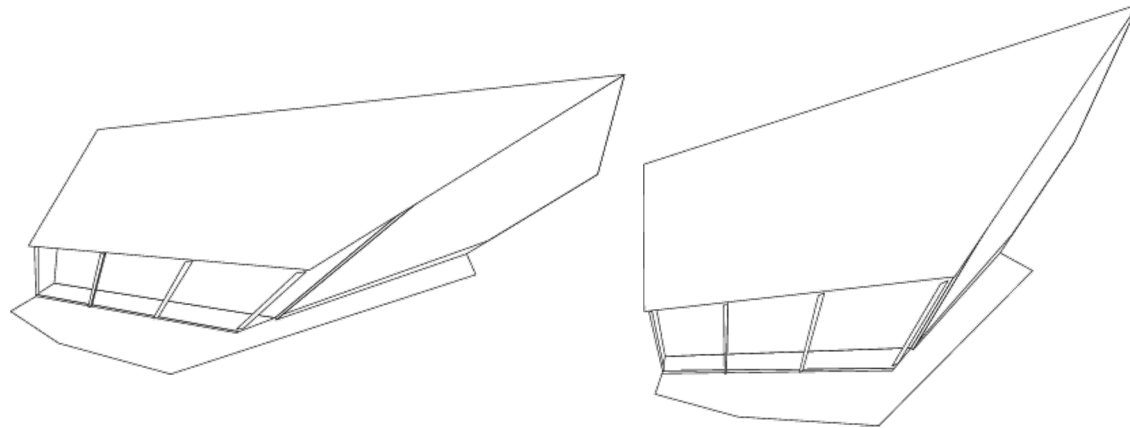
- With no constraints on the camera calibration matrices or on the scene, we can reconstruct up to a *projective* ambiguity:

$$x \cong PX = (PQ^{-1})(QX)$$

Q is a general full-rank 4×4 matrix



Projective ambiguity



Affine ambiguity

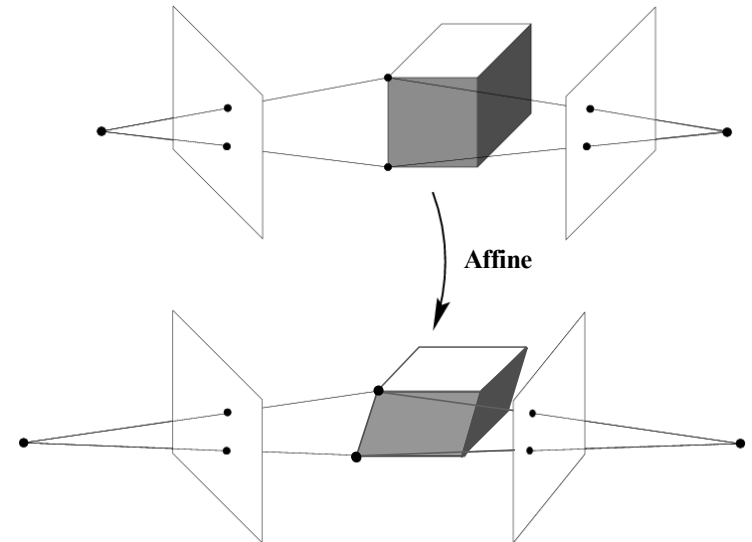
- If we impose parallelism constraints, we can get a reconstruction up to an *affine* ambiguity:

$$x \cong PX = (PQ_A^{-1})(Q_A X)$$

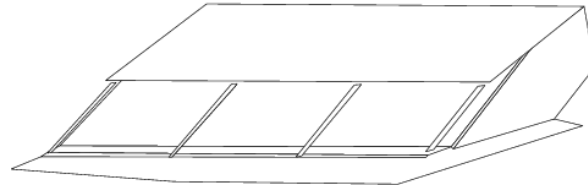
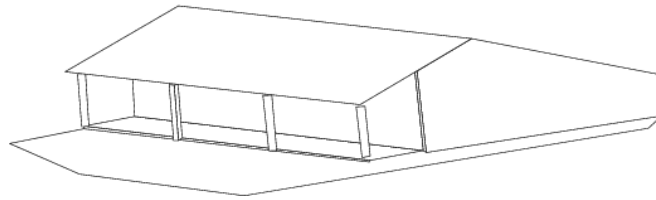
3×3 full-rank matrix

3×1 translation vector

$$Q_A = \begin{bmatrix} A & t \\ \mathbf{0}^T & 1 \end{bmatrix}$$



Affine ambiguity



Similarity ambiguity

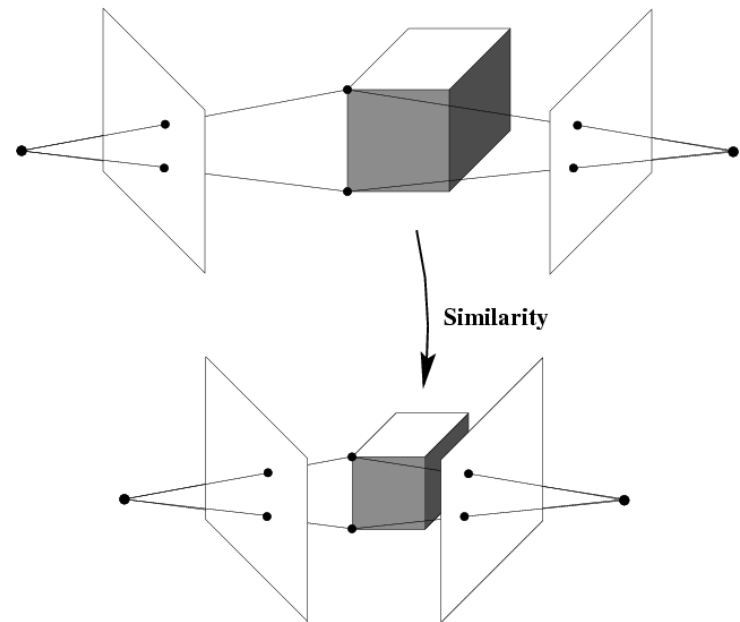
- A reconstruction that obeys orthogonality constraints on camera parameters and/or scene

$$x \cong PX = (PQ_S^{-1})(Q_S X)$$

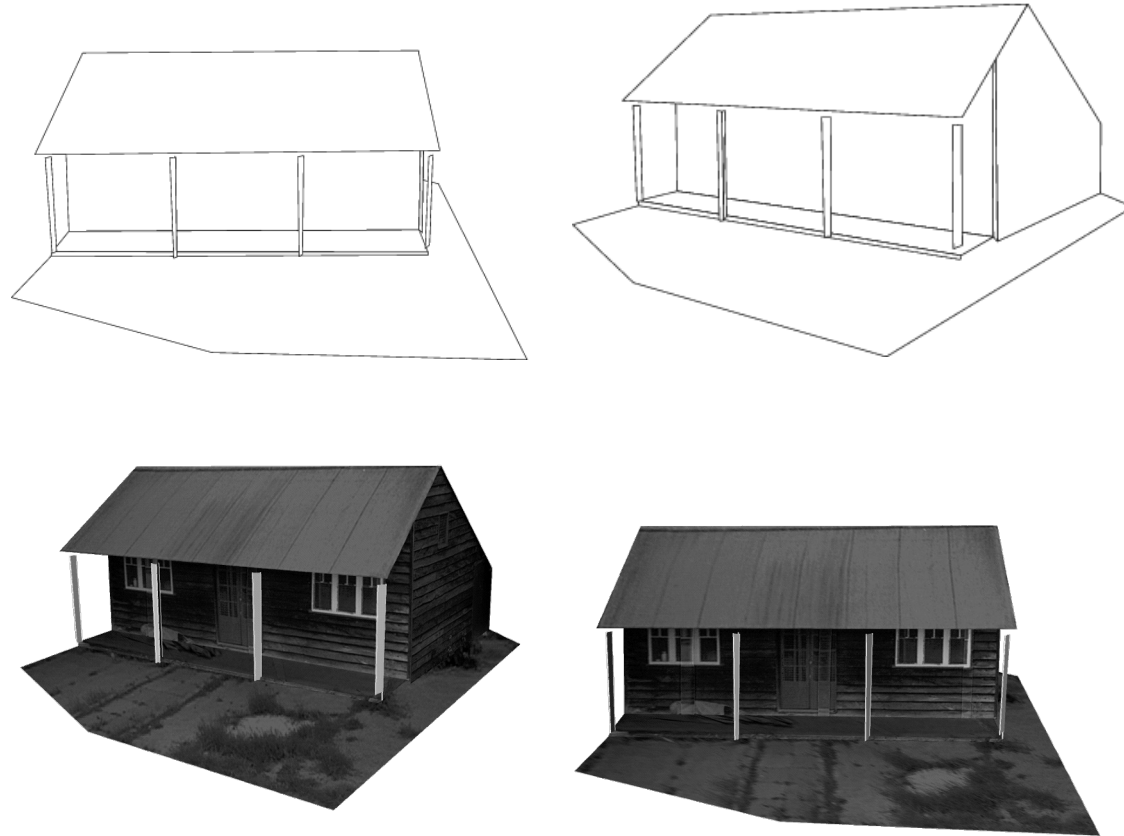
3×3 rotation matrix

3×1 translation vector

$$Q_S = \begin{bmatrix} sR & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$$



Similarity ambiguity

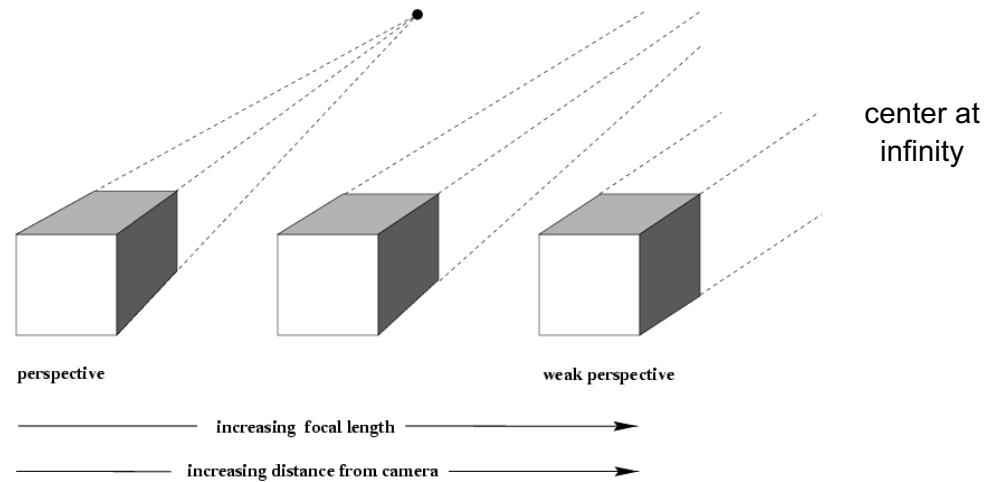
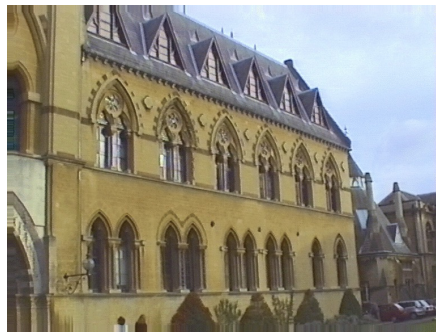


Outline: Structure from motion

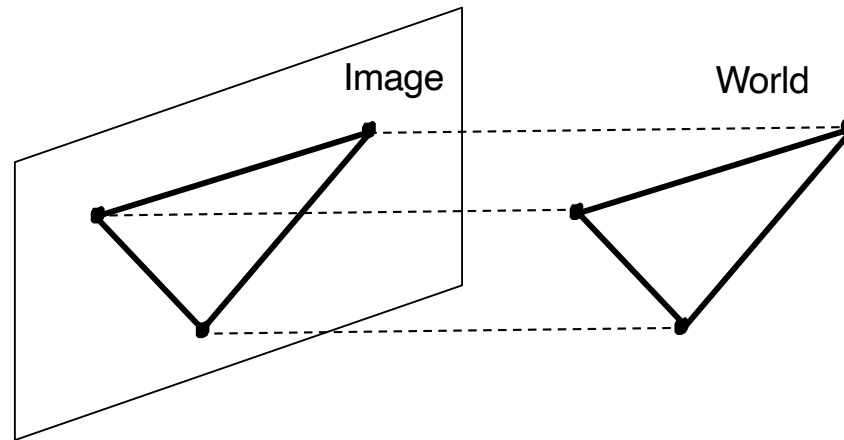
- Problem definition and ambiguities
- Affine structure from motion
 - Factorization

Affine structure from motion

- Let's start with *affine* or *weak perspective* cameras



Recall: Orthographic projection



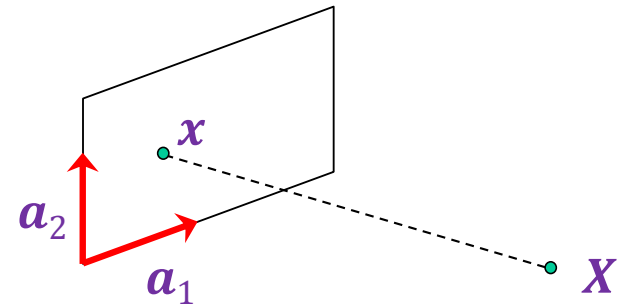
Just drop the z coordinate!

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

General affine projection

- A general affine projection is a 3D-to-2D linear mapping plus translation:

$$P = \begin{bmatrix} a_{11} & a_{12} & a_{13} & t_1 \\ a_{21} & a_{22} & a_{23} & t_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} A & t \\ \mathbf{0}^T & 1 \end{bmatrix}$$



a_1, a_2 : rows of projection matrix

- In non-homogeneous coordinates:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = AX + t$$

Projection of
world origin

Affine structure from motion

- **Given:** m images of n fixed 3D points such that

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{X}_j + \mathbf{t}_i, \quad i = 1, \dots, m, j = 1, \dots, n$$

- **Problem:** use the mn correspondences \mathbf{x}_{ij} to estimate m projection matrices \mathbf{A}_i and translation vectors \mathbf{t}_i , and n points \mathbf{X}_j
- The reconstruction is defined up to an arbitrary *affine* transformation \mathbf{Q} (12 degrees of freedom):

$$\begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \mathbf{Q}^{-1}, \quad \begin{pmatrix} \mathbf{X}_j \\ 1 \end{pmatrix} \rightarrow \mathbf{Q} \begin{pmatrix} \mathbf{X}_j \\ 1 \end{pmatrix}$$

- How many knowns and unknowns for m images and n points?
 - $2mn$ knowns and $8m + 3n$ unknowns
 - To be able to solve this problem, we must have $2mn \geq 8m + 3n - 12$ (affine ambiguity takes away 12 dof)
 - E.g., for **two** views, we need **four** point correspondences

Affine structure from motion

- First, center the data by subtracting the centroid of the image points in each view:

$$\begin{aligned}\hat{x}_{ij} &= x_{ij} - \frac{1}{n} \sum_{k=1}^n x_{ik} \\ &= \mathbf{A}_i \mathbf{X}_j + \mathbf{t}_i - \frac{1}{n} \sum_{k=1}^n (\mathbf{A}_i \mathbf{X}_k + \mathbf{t}_i) \\ &= \mathbf{A}_i \left(\mathbf{X}_j - \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k \right) \\ &= \mathbf{A}_i \hat{\mathbf{X}}_j\end{aligned}$$

Affine structure from motion

- After centering, each normalized 2D point \hat{x}_{ij} is related to the 3D point by

$$\hat{x}_{ij} = A_i \hat{X}_j$$

- We can get rid of the need to center the 3D data (and the translation ambiguity) by defining the origin of the world coordinate system as the centroid of the 3D points

Affine structure from motion

- Let's create a $2m \times n$ data (measurement) matrix:

$$D = \begin{bmatrix} \hat{\mathbf{x}}_{11} & \hat{\mathbf{x}}_{12} & \cdots & \hat{\mathbf{x}}_{1n} \\ \hat{\mathbf{x}}_{21} & \hat{\mathbf{x}}_{22} & \cdots & \hat{\mathbf{x}}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{x}}_{m1} & \hat{\mathbf{x}}_{m2} & \cdots & \hat{\mathbf{x}}_{mn} \end{bmatrix}$$

↓ cameras (2m)

→ points (n)

$$\hat{\mathbf{x}}_{ij} = A_i \mathbf{X}_j$$

Affine structure from motion

- Let's create a $2m \times n$ data (measurement) matrix:

$$D = \begin{bmatrix} \hat{x}_{11} & \hat{x}_{12} & \cdots & \hat{x}_{1n} \\ \hat{x}_{21} & \hat{x}_{22} & \cdots & \hat{x}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}_{m1} & \hat{x}_{m2} & \cdots & \hat{x}_{mn} \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}$$

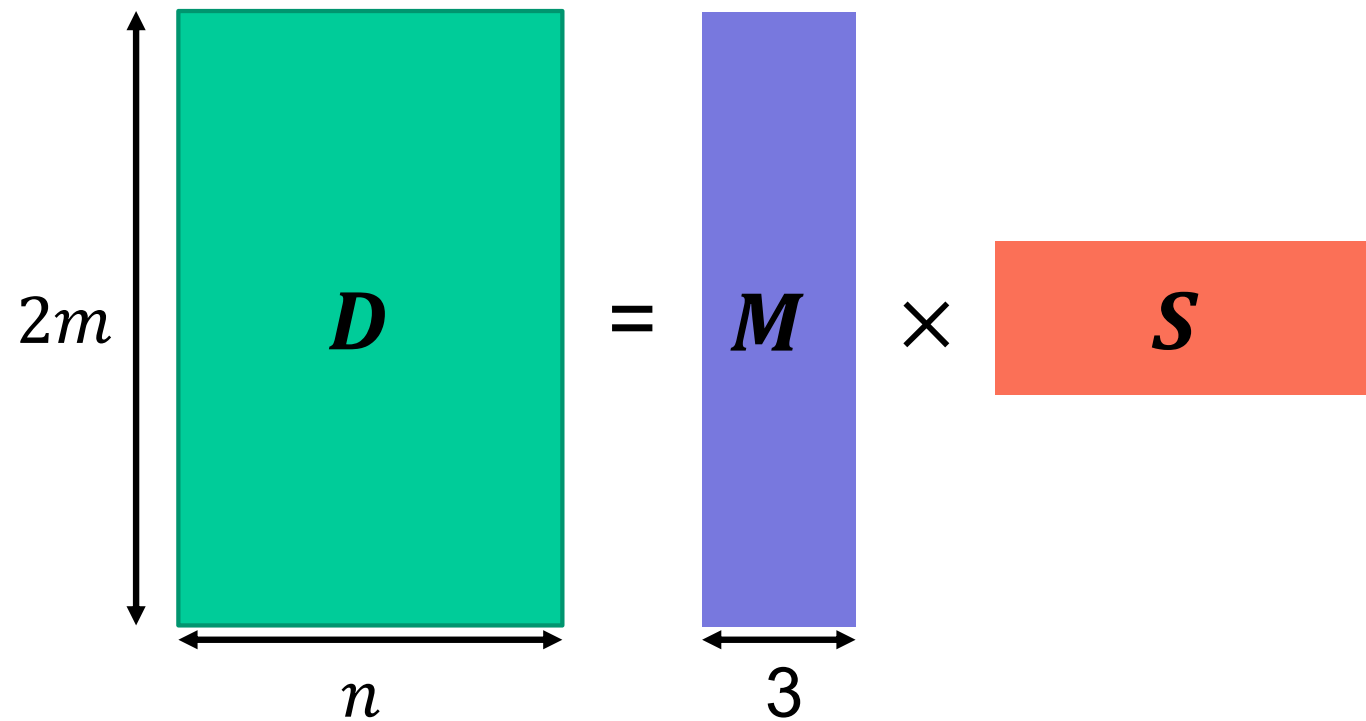
M
cameras
($2m \times 3$)

S
points ($3 \times n$)

- What must be the rank of the measurement matrix $D = MS$?

Factorizing the measurement matrix

- We want:



Factorizing the measurement matrix

- Perform SVD of D :

The diagram illustrates the SVD factorization of matrix D . It shows a green rectangle representing matrix D with dimensions $2m \times n$. This is followed by an equals sign, then a purple rectangle representing matrix U with dimensions $2m \times n$. This is followed by a multiplication sign, then a red square representing matrix Σ with dimensions $n \times n$. This is followed by another multiplication sign, then another red square representing matrix V^T with dimensions $n \times n$.

$$\begin{matrix} \mathbf{D} \\ 2m \times n \end{matrix} = \begin{matrix} \mathbf{U} \\ 2m \times n \end{matrix} \times \begin{matrix} \mathbf{\Sigma} \\ n \times n \end{matrix} \times \begin{matrix} \mathbf{V}^T \\ n \times n \end{matrix}$$

Factorizing the measurement matrix

- Keep top 3 singular values:

- This is the closest approximation of D with a rank-3 matrix in terms of Frobenius norm

$$\begin{matrix} D \\ 2m \times n \end{matrix} = \begin{matrix} U_3 \\ 2m \times 3 \end{matrix} \times \begin{matrix} \Sigma_3 \\ 3 \times 3 \end{matrix} \times \begin{matrix} V_3^T \\ 3 \times n \end{matrix}$$

- What to do about Σ_3 ?

- One solution: $M = U_3 \Sigma_3^{\frac{1}{2}}, S = \Sigma_3^{\frac{1}{2}} V_3^T$

Factorizing the measurement matrix

- One possible solution:

$$\begin{matrix} \mathbf{D} \\ 2m \times n \end{matrix} = \begin{matrix} \mathbf{M} \\ 2m \times 3 \end{matrix} \times \begin{matrix} \mathbf{S} \\ 3 \times n \end{matrix}$$

$$\mathbf{S} = \Sigma_3^{\frac{1}{2}} \mathbf{V}_3^T$$

$$\mathbf{M} = \mathbf{U}_3 \Sigma_3^{\frac{1}{2}}$$

- Are there other solutions?

Factorizing the measurement matrix

- Other possible solutions:

The diagram illustrates the factorization of the measurement matrix D into four components: M , Q , Q^{-1} , and S . Each component is represented by a colored rectangle with its dimensions and name inside. D is a green rectangle with dimensions $2m \times n$. M is a purple rectangle with dimensions $2m \times 3$. Q is a pink square with dimensions 3×3 . Q^{-1} is a pink square with dimensions 3×3 . S is an orange rectangle with dimensions $3 \times n$. The equation is $D = M \times Q \times Q^{-1} \times S$.

$$\begin{matrix} \mathbf{D} \\ 2m \times n \end{matrix} = \begin{matrix} \mathbf{M} \\ 2m \times 3 \end{matrix} \times \begin{matrix} \mathbf{Q} \\ 3 \times 3 \end{matrix} \times \begin{matrix} \mathbf{Q}^{-1} \\ 3 \times 3 \end{matrix} \times \begin{matrix} \mathbf{S} \\ 3 \times n \end{matrix}$$

We can estimate Q to give the camera matrices in M desirable properties, like orthographic projection

Eliminating the affine ambiguity

- So far, we have obtained one solution:

$$D = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}$$

$2m \times 3$ $3 \times n$

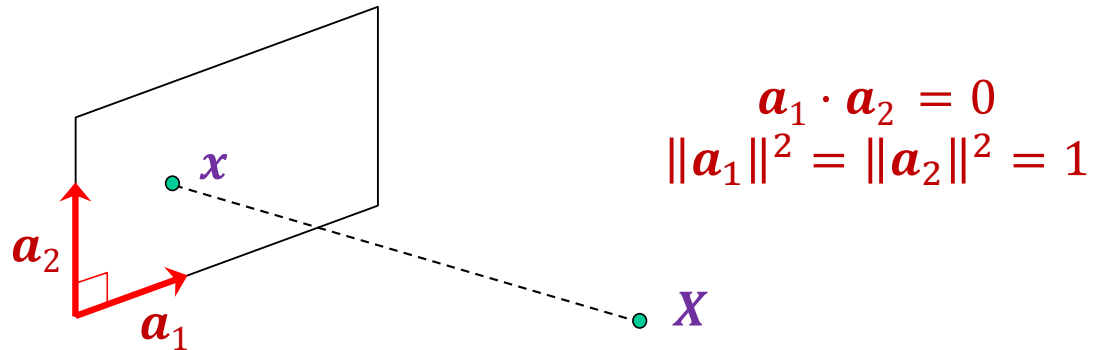
- We want:

$$D = \begin{bmatrix} A_1 Q \\ A_2 Q \\ \vdots \\ A_m Q \end{bmatrix} \begin{bmatrix} Q^{-1} X_1 & Q^{-1} X_2 & \cdots & Q^{-1} X_n \end{bmatrix}$$

such that each camera matrix $A_i Q$ represents orthographic projection, i.e., has orthonormal axes (rows)

Eliminating the affine ambiguity

- Let \mathbf{a}_1 and \mathbf{a}_2 be the rows of a 2×3 orthographic projection matrix. Then



- This translates into $3m$ constraints on the 9 entries of \mathbf{Q} :

$$(\mathbf{A}_i \mathbf{Q})(\mathbf{A}_i \mathbf{Q})^T = \mathbf{A}_i (\mathbf{Q} \mathbf{Q}^T) \mathbf{A}_i^T = \mathbf{I}_{2 \times 2}, \quad i = 1, \dots, m$$

- Are the constraints linear?
- First, solve for $\mathbf{L} = \mathbf{Q} \mathbf{Q}^T$
- Recover \mathbf{Q} from \mathbf{L} by Cholesky decomposition
- Update \mathbf{M} to $\mathbf{M} \mathbf{Q}$, \mathbf{S} to $\mathbf{Q}^{-1} \mathbf{S}$

Reconstruction results



1



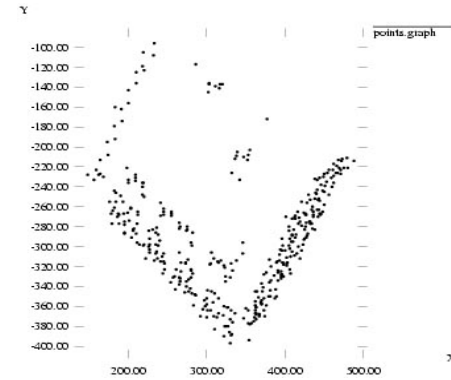
60



120



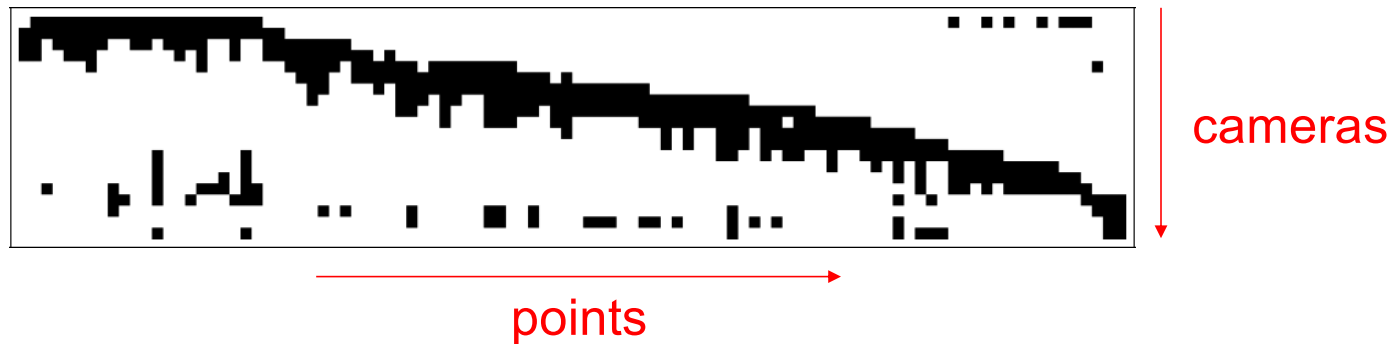
150



C. Tomasi and T. Kanade, [Shape and motion from image streams under orthography: A factorization method](#), IJCV 1992

Dealing with missing data

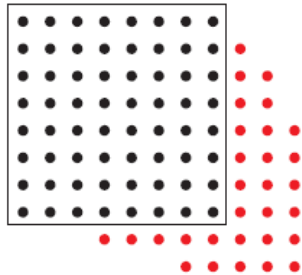
- So far, we have assumed that all points are visible in all views
- In reality, the measurement matrix typically looks something like this:



- Possible solution: decompose matrix into dense sub-blocks, factorize each sub-block, and fuse the results
 - Unfortunately, finding dense maximal sub-blocks of the matrix is NP-complete (equivalent to finding maximal cliques in a graph)

Dealing with missing data

- Incremental bilinear refinement:



Perform factorization
on a dense sub-block

Solve for a new 3D
point visible by at
least two known
cameras –
triangulation

Solve for a new
camera that sees at
least three known 3D
points – **calibration**

Outline: Structure from motion

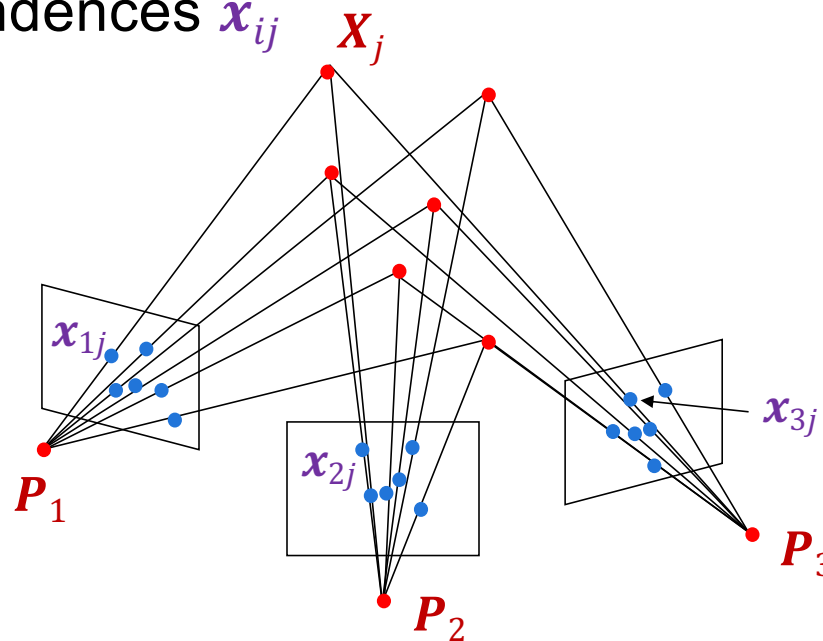
- Problem definition and ambiguities
- Affine structure from motion
 - Factorization
- Projective structure from motion

Projective structure from motion

- **Given:** m images of n fixed 3D points such that (ignoring visibility):

$$\mathbf{x}_{ij} \cong \mathbf{P}_i \mathbf{X}_j, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

- **Problem:** estimate m projection matrices \mathbf{P}_i and n 3D points \mathbf{X}_j from the mn correspondences \mathbf{x}_{ij}



Projective structure from motion

- **Given:** m images of n fixed 3D points such that (ignoring visibility):

$$\mathbf{x}_{ij} \cong \mathbf{P}_i \mathbf{X}_j, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

- **Problem:** estimate m projection matrices \mathbf{P}_i and n 3D points \mathbf{X}_j from the mn correspondences \mathbf{x}_{ij}
- With no calibration info, cameras and points can only be recovered up to a 4×4 projective transformation \mathbf{Q} :

$$\mathbf{X} \rightarrow \mathbf{Q}\mathbf{X}, \mathbf{P} \rightarrow \mathbf{P}\mathbf{Q}^{-1}$$

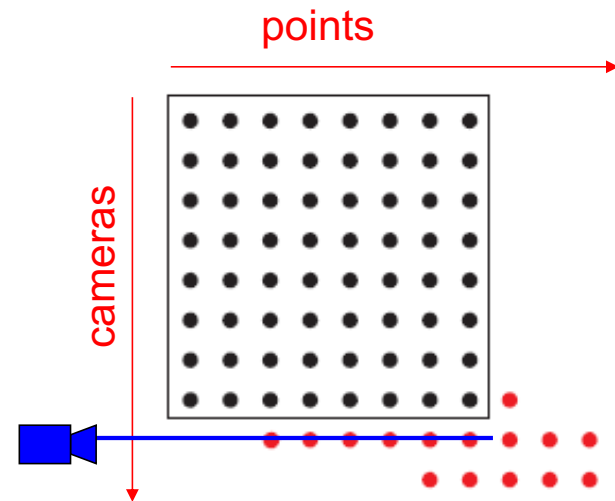
- We can solve for structure and motion when $2mn \geq 11m + 3n - 15$
- For two cameras, at least **7 points** are needed

Projective SFM: Two-camera case

1. Estimate fundamental matrix F between the two views
 2. Set first camera matrix to $[I \mid \mathbf{0}]$
 3. Then the second camera matrix is given by $[A \mid t]$ where t is the epipole ($F^T t = \mathbf{0}$) and $A = -[t_{\times}]F$
- In practice, SFM pipelines use guesses of intrinsic parameters and the [five-point algorithm](#)

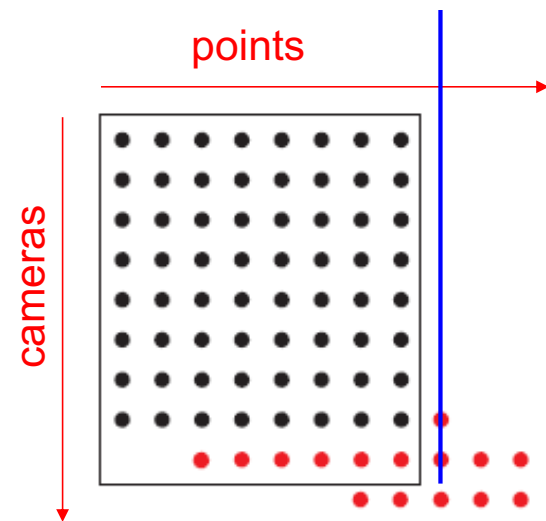
Incremental structure from motion

- Initialize motion from two images using fundamental matrix
- Initialize structure by triangulation
- For each additional view:
 - Determine projection matrix of new camera using all the known 3D points that are visible in its image – **calibration**



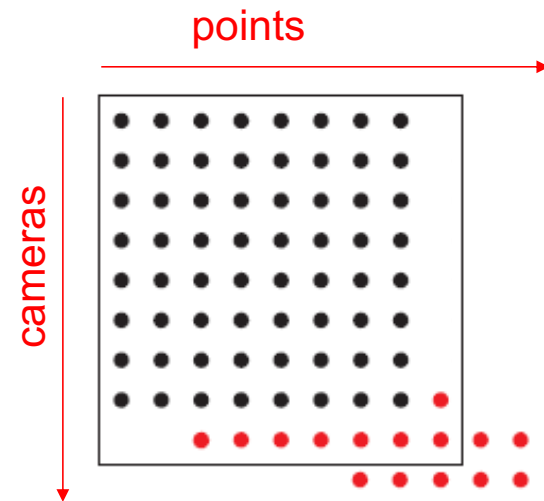
Incremental structure from motion

- Initialize motion from two images using fundamental matrix
- Initialize structure by triangulation
- For each additional view:
 - Determine projection matrix of new camera using all the known 3D points that are visible in its image – **calibration**
 - Refine and extend structure: compute new 3D points, re-optimize existing points that are also seen by this camera – **triangulation**



Incremental structure from motion

- Initialize motion from two images using fundamental matrix
- Initialize structure by triangulation
- For each additional view:
 - Determine projection matrix of new camera using all the known 3D points that are visible in its image – **calibration**
 - Refine and extend structure: compute new 3D points, re-optimize existing points that are also seen by this camera – **triangulation**
- Refine structure and motion: bundle adjustment

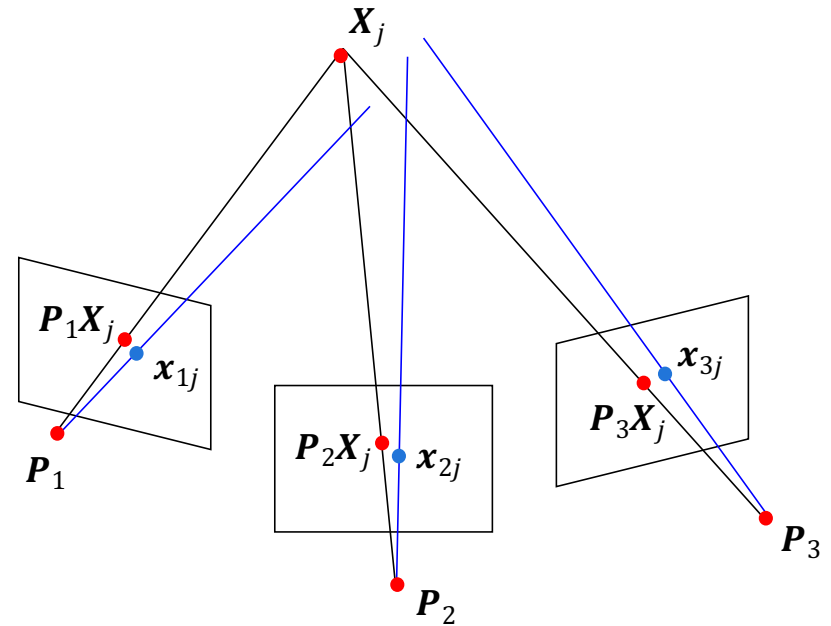


Bundle adjustment

- Non-linear method for refining structure and motion
- Minimize reprojection error (with lots of bells and whistles):

$$\sum_{i=1}^m \sum_{j=1}^n w_{ij} d \left(\mathbf{x}_{ij} - \text{proj}(\mathbf{P}_i \mathbf{X}_j) \right)^2$$

visibility flag: is
point j visible in
view i ?



Outline: Structure from motion

- Problem definition and ambiguities
- Affine structure from motion
 - Factorization
- Projective structure from motion
 - Incremental reconstruction, bundle adjustment
- **Modern structure from motion pipeline**

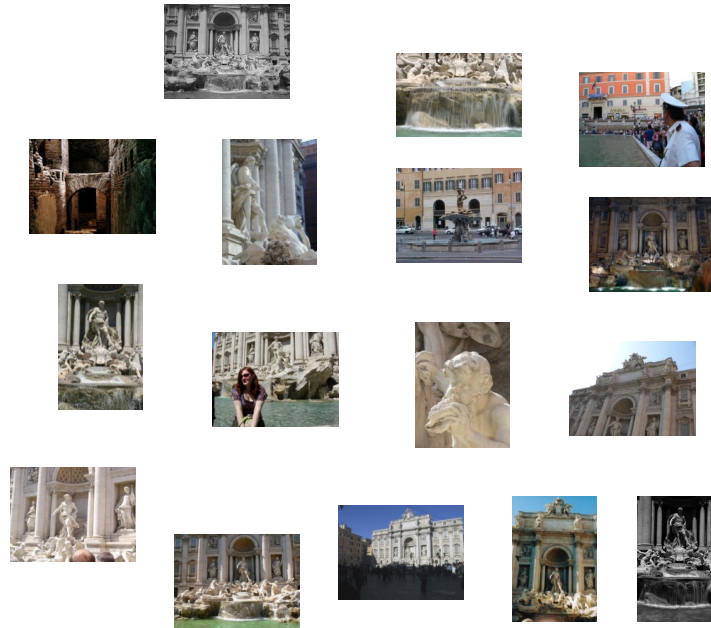
Representative SFM pipeline



N. Snavely, S. Seitz, and R. Szeliski. [Photo tourism: Exploring photo collections in 3D](http://phototour.cs.washington.edu/). SIGGRAPH 2006
<http://phototour.cs.washington.edu/>

Feature detection

Detect SIFT features

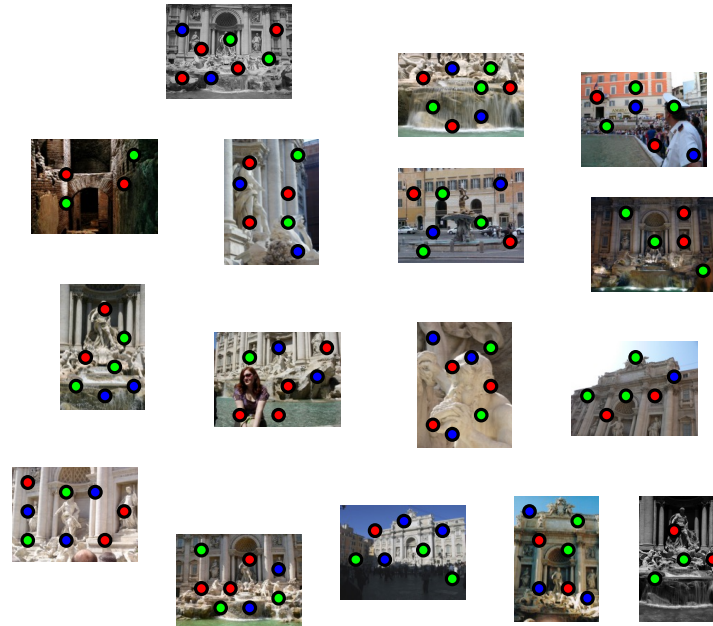


Source: N. Snavely

Feature detection

Detect SIFT features

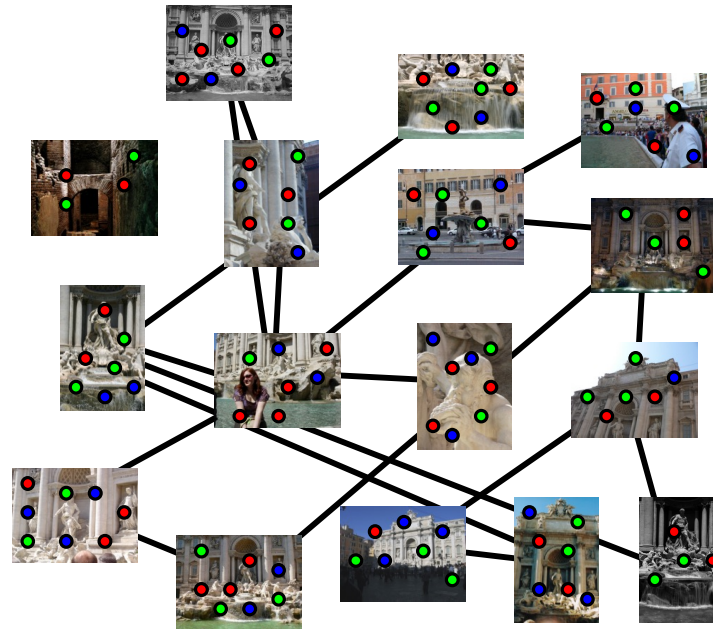
Other popular feature types: [SURF](#), [ORB](#), [BRISK](#), ...



Source: N. Snavely

Feature matching

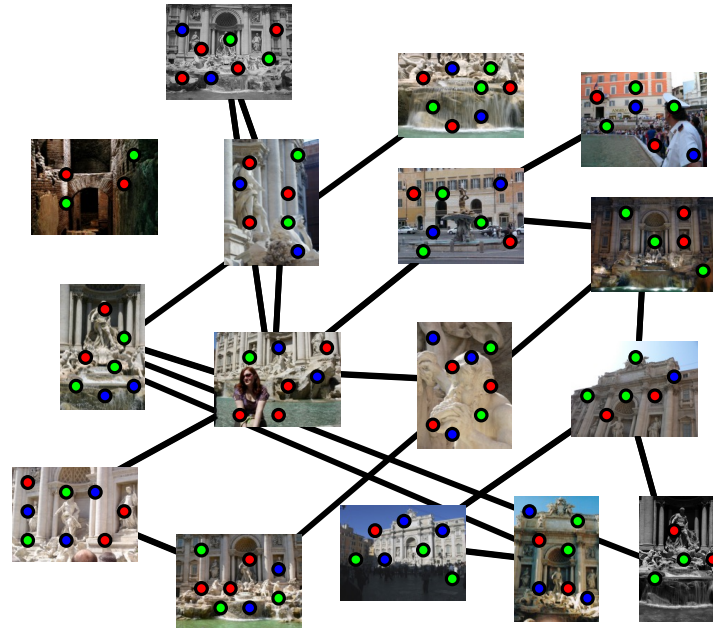
Match features between each pair of images



Source: N. Snavely

Feature matching

Use RANSAC to estimate fundamental matrix between each pair



Source: N. Snavely

Feature matching

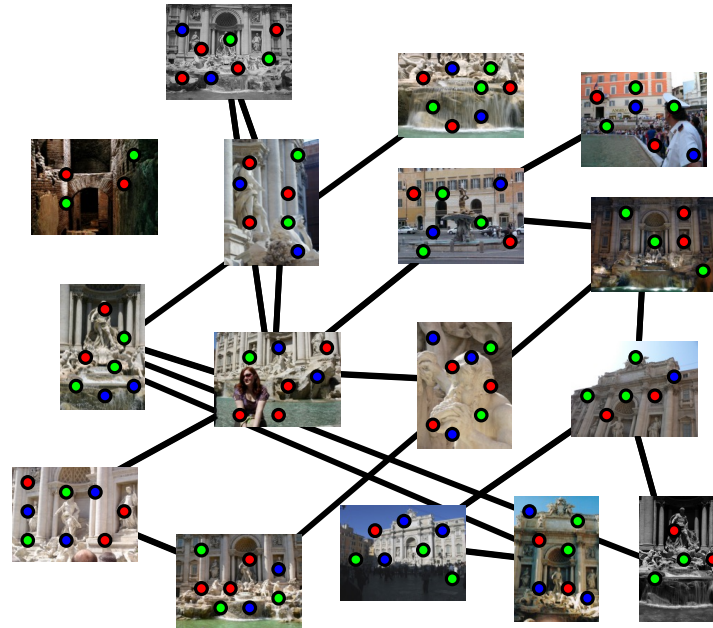
Use RANSAC to estimate fundamental matrix between each pair



[Image source](#)

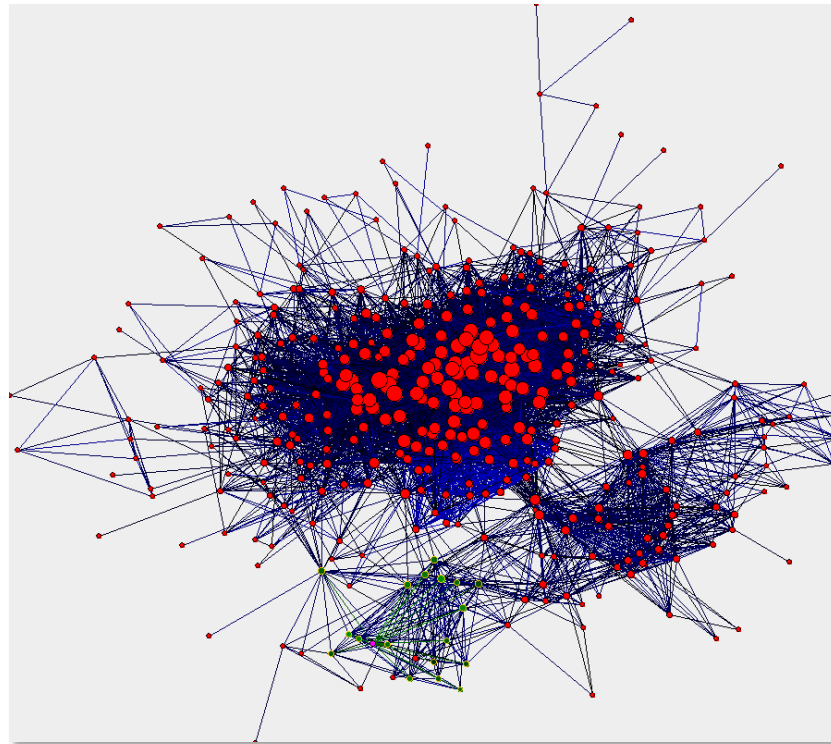
Feature matching

Use RANSAC to estimate fundamental matrix between each pair



Source: N. Snavely

Image connectivity graph



(graph layout produced using the Graphviz toolkit: <http://www.graphviz.org/>)

Source: N. Snavely

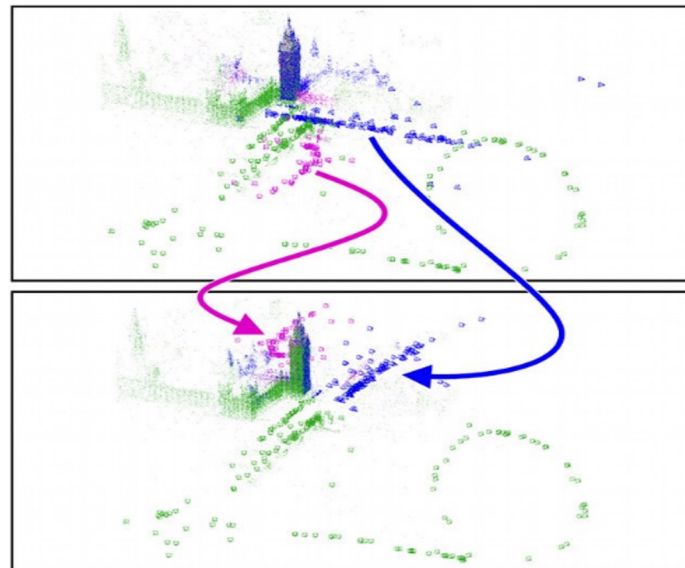
Incremental SFM

- Pick a pair of images with lots of inliers (and preferably, good EXIF data)
 - Initialize intrinsic parameters (focal length, principal point) from EXIF
 - Estimate extrinsic parameters (R and t) using [five-point algorithm](#)
 - Use triangulation to initialize model points
- While remaining images exist
 - Find an image with many feature matches with images in the model
 - Run RANSAC on feature matches to register new image to model
 - Triangulate new points
 - Perform bundle adjustment to re-optimize everything
 - Optionally, align with GPS from EXIF data or ground control points

The devil is in the details

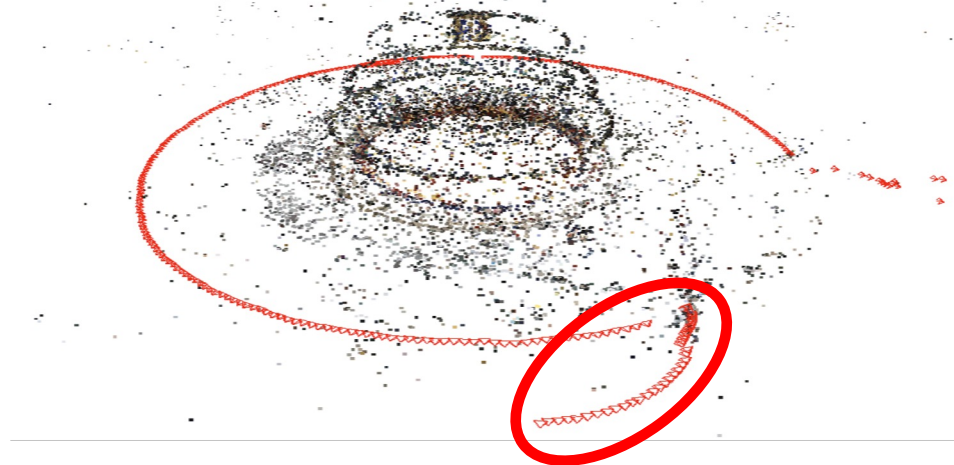
- Handling degenerate configurations (e.g., homographies)
- Filtering out incorrect matches
- Dealing with repetitions and symmetries

Repetitive structures cause catastrophic failures



<https://demuc.de/tutorials/cvpr2017/sparse-modeling.pdf>

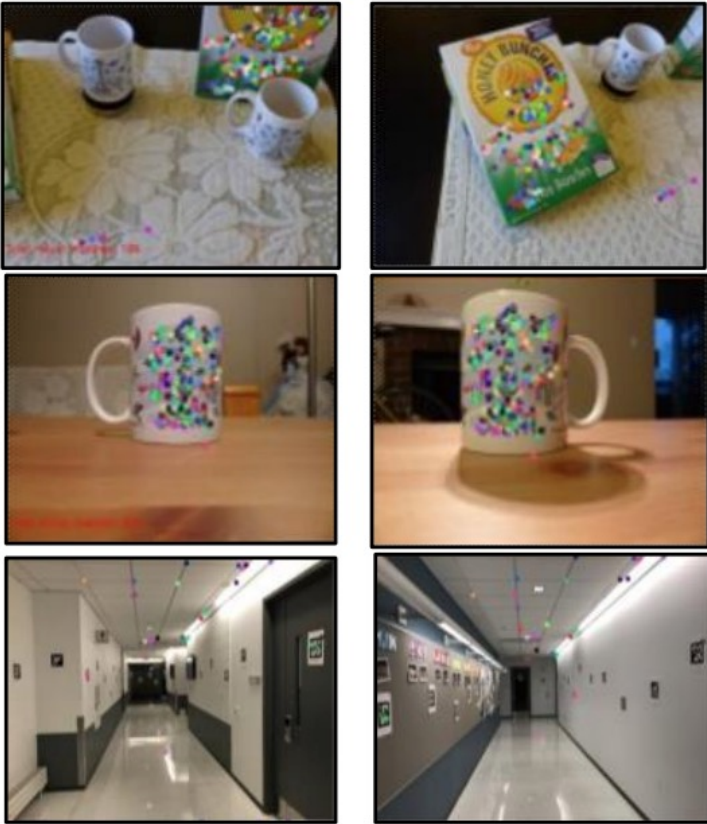
Repetitive structures cause catastrophic failures



R. Kataria et al. [Improving Structure from Motion with Reliable Resectioning](#). 3DV 2020

Repetitive structures cause catastrophic failures

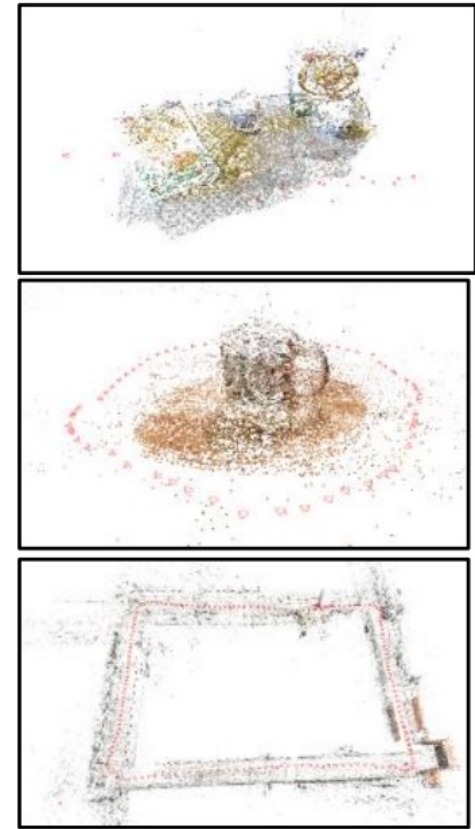
Erroneously Matching Images



Baseline Reconstruction



Our Reconstruction



The devil is in the details

- Handling degenerate configurations (e.g., homographies)
- Filtering out incorrect matches
- Dealing with repetitions and symmetries
- Reducing error accumulation and closing loops

Reducing error accumulation and closing loops



seattle1

more_half

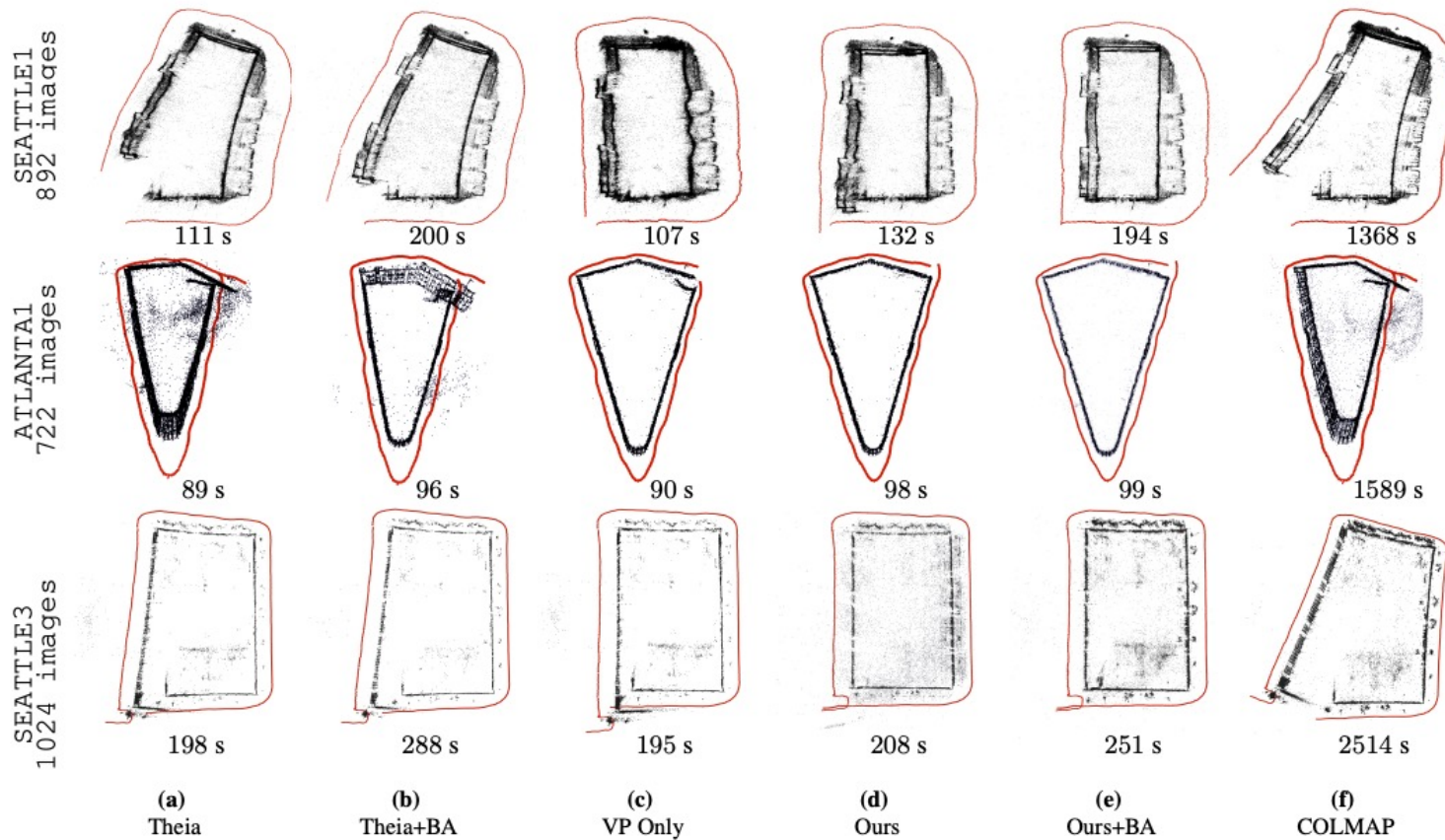
seattle2

atlanta1

seattle3

A. Holynski et al. [Reducing Drift in Structure From Motion Using Extended Features](#). arXiv 2020

Reducing error accumulation and closing loops



A. Holynski et al. [Reducing Drift in Structure From Motion Using Extended Features](#). arXiv 2020

The devil is in the details

- Handling degenerate configurations (e.g., homographies)
- Filtering out incorrect matches
- Dealing with repetitions and symmetries
- Reducing error accumulation and closing loops
- Making the whole thing efficient!
 - See, e.g., [Towards Linear-Time Incremental Structure from Motion](#)

SFM software

- [Bundler](#)
- [OpenSfM](#)
- [OpenMVG](#)
- [VisualSFM](#)
- [COLMAP](#)
- See also [Wikipedia's list of toolboxes](#)