CS4670/5670: Computer Vision

Kavita Bala

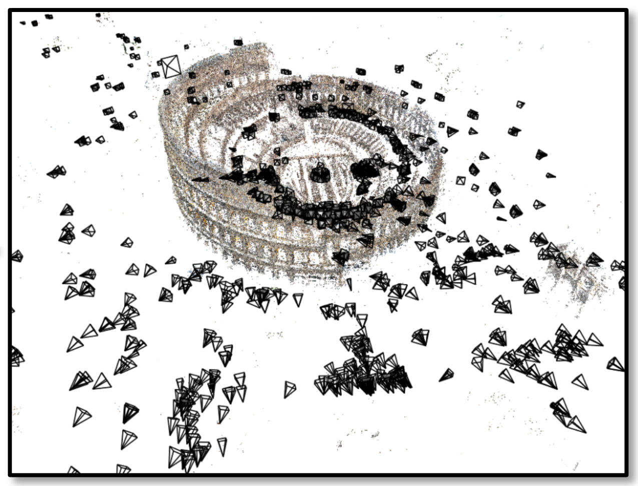
Lecture 25: Structure from motion

Announcements

- HW 2 out.
  - New version from yesterday with some clarifications on coordinate systems
- PA 4 tonight

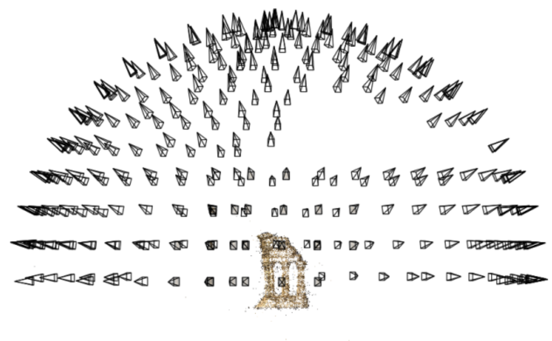
Structure from motion

- Given many images, how can we
  a) figure out where they were all taken from?
  b) build a 3D model of the scene?

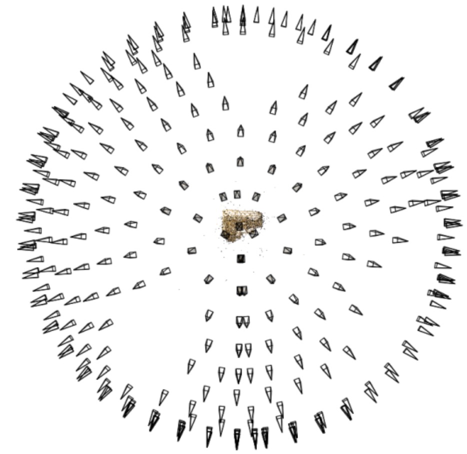


This is (roughly) the **structure from motion** problem

Structure from motion



Reconstruction (side)



(top)

- Input: images with points in correspondence
  $p_{i,j} = (u_{i,j}, v_{i,j})$

- Output
  - structure: 3D location $\mathbf{x}_i$ for each point p_i
  - motion: camera parameters $\mathbf{R}_j$, $\mathbf{t}_j$ possibly $\mathbf{K}_j$

- Objective function: minimize *reprojection error*

What we've seen so far…

- 2D transformations between images
  - Translations, affine transformations, homographies…
- Fundamental matrices
  - Still represent relationships between 2D images
- **What's new:** Explicitly representing 3D geometry of cameras *and points*

Camera calibration and triangulation

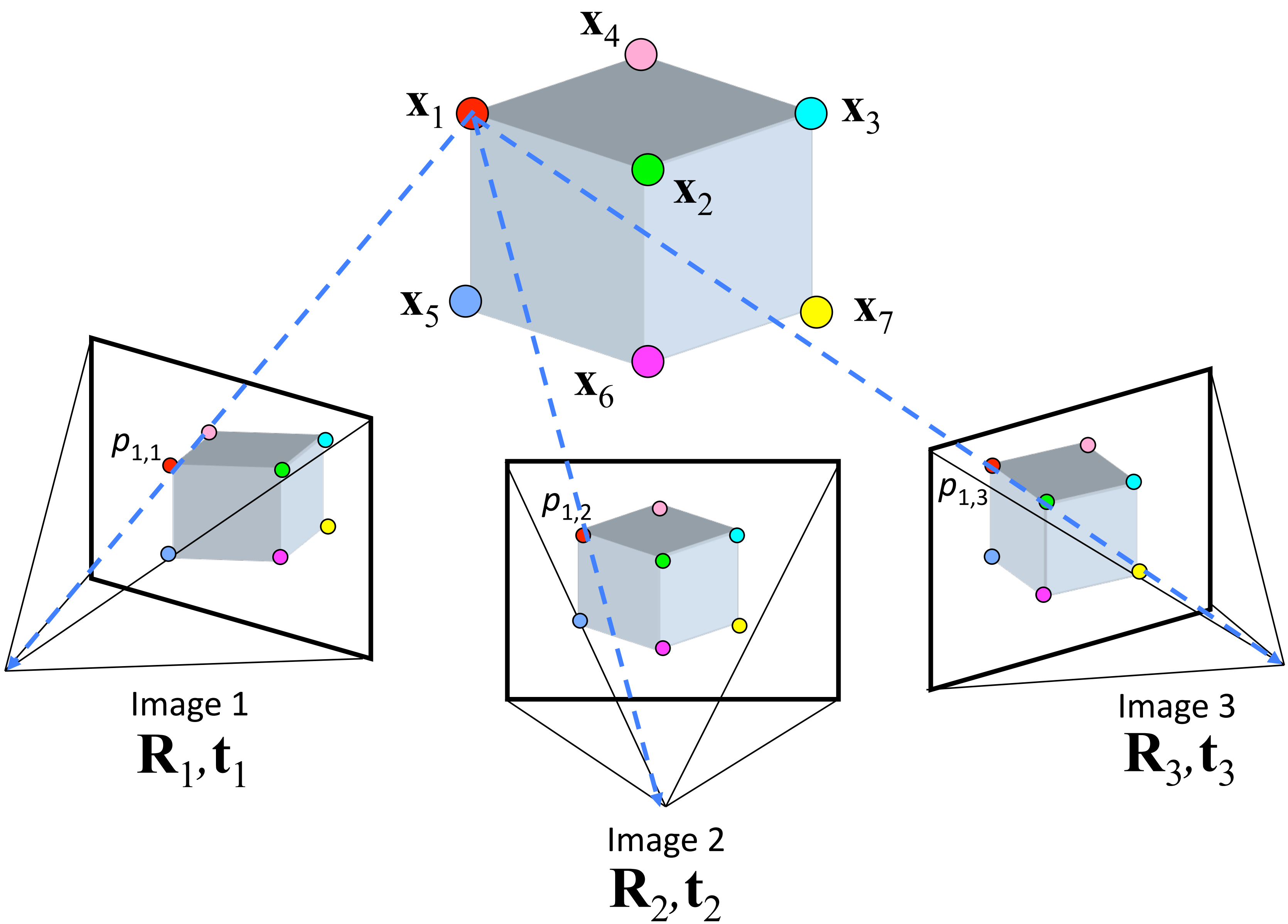
- Suppose we know 3D points
  - And have matches between these points and an image
  - How can we compute the camera parameters?

- Suppose we have know camera parameters, each of which observes a point
  - How can we compute the 3D location of that point?

Structure from motion

- SfM solves both of these problems *at once*
- A kind of chicken-and-egg problem
  - (but solvable)

Also doable from video



$\mathbf{x}_4$
$\mathbf{x}_1$
$\mathbf{x}_3$
$\mathbf{x}_2$
$\mathbf{x}_5$
$\mathbf{x}_7$
$\mathbf{x}_6$
$p_{1,1}$
$p_{1,2}$
$p_{1,3}$
Image 1
$\mathbf{R}_1, \mathbf{t}_1$
Image 2
$\mathbf{R}_2, \mathbf{t}_2$
Image 3
$\mathbf{R}_3, \mathbf{t}_3$

Structure from motion

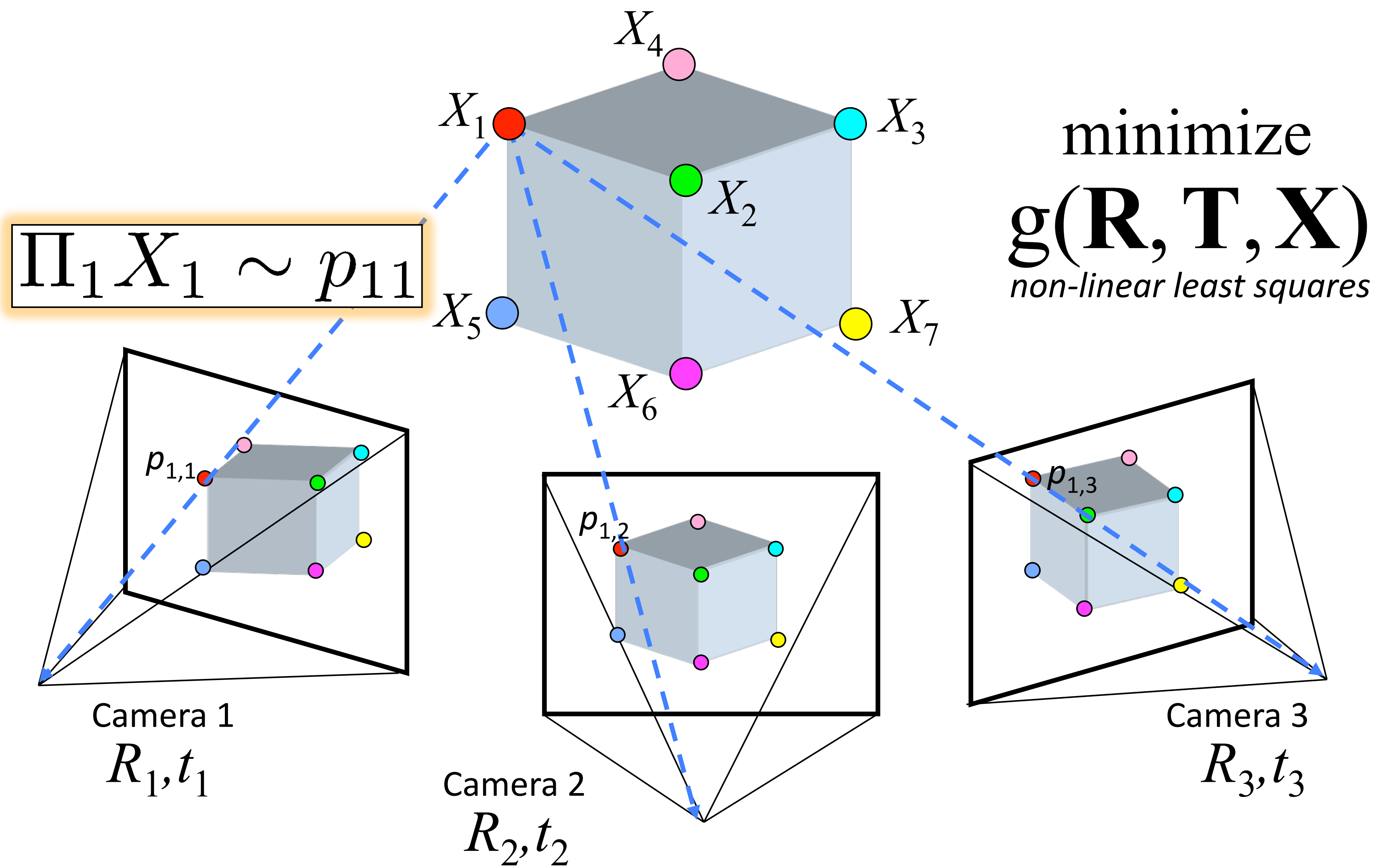
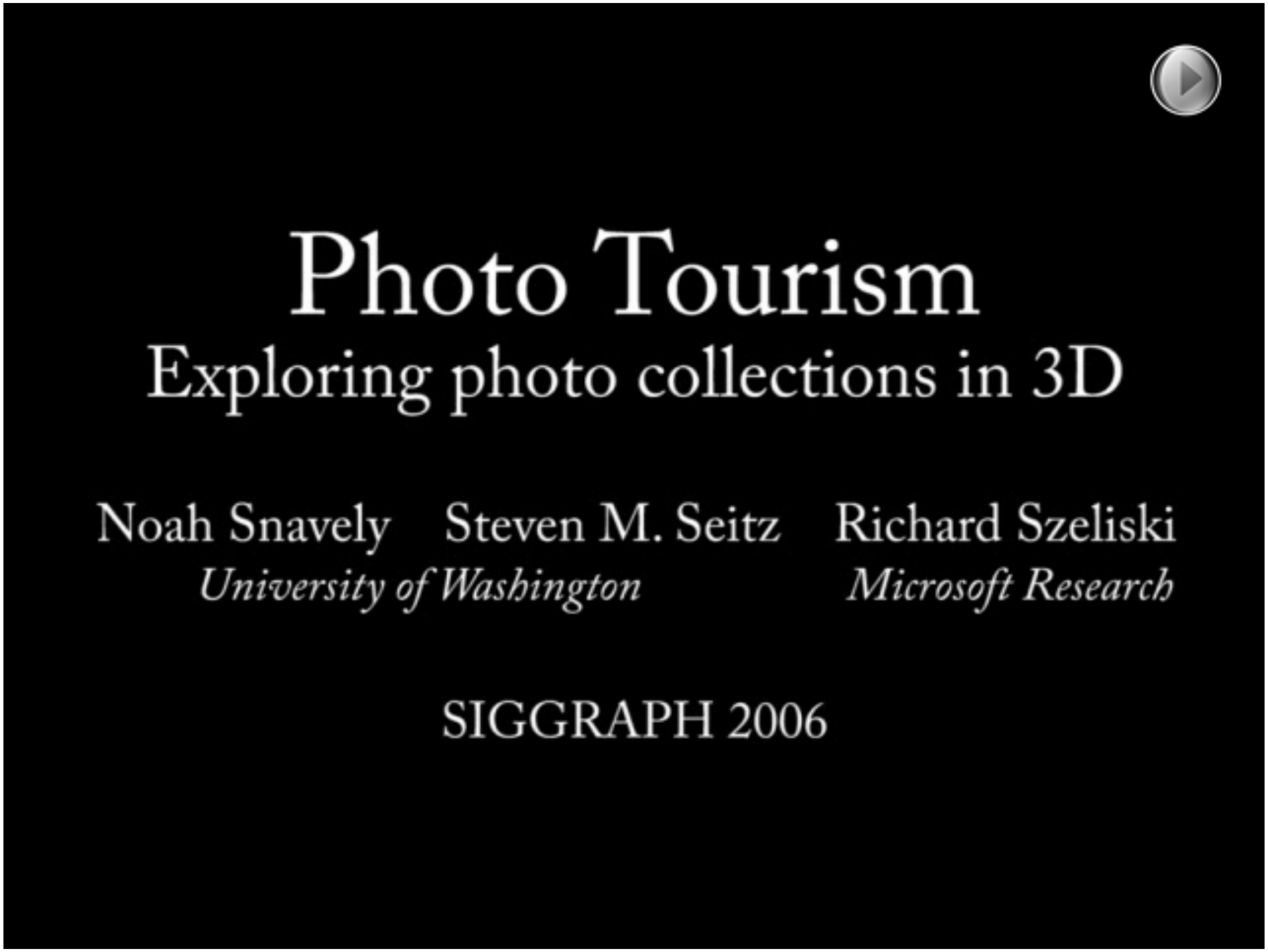
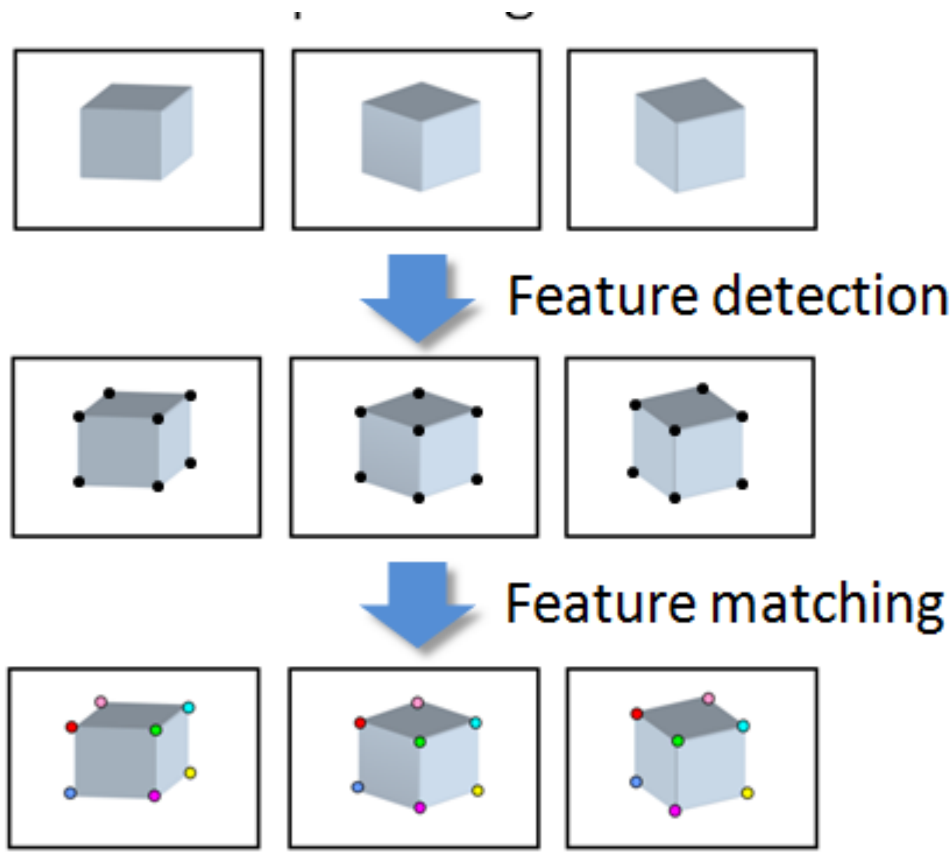


Photo Tourism



Input

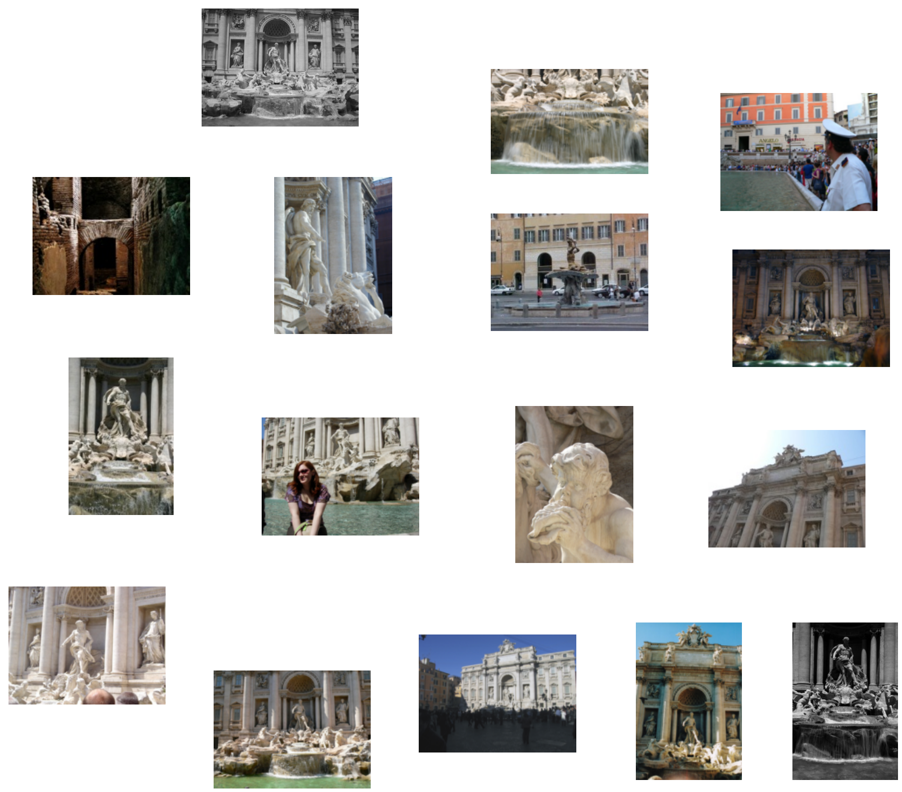


First step: how to get correspondence?

- Feature detection and matching

Feature detection

Detect features using SIFT [Lowe, IJCV 2004]

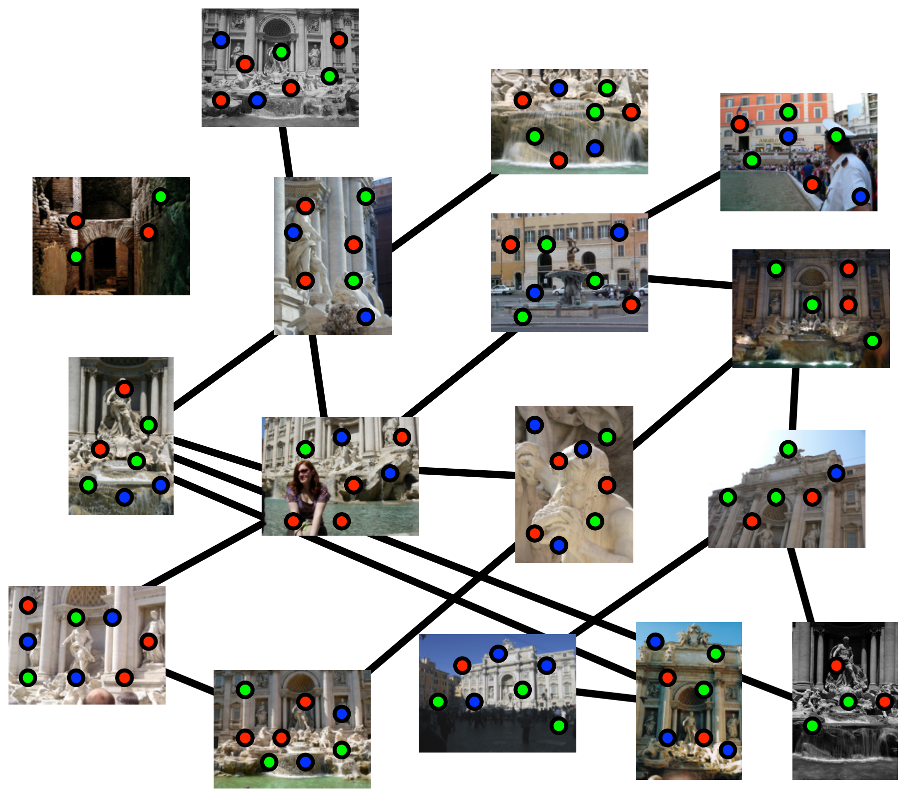


Feature detection

Detect features using SIFT [Lowe, IJCV 2004]

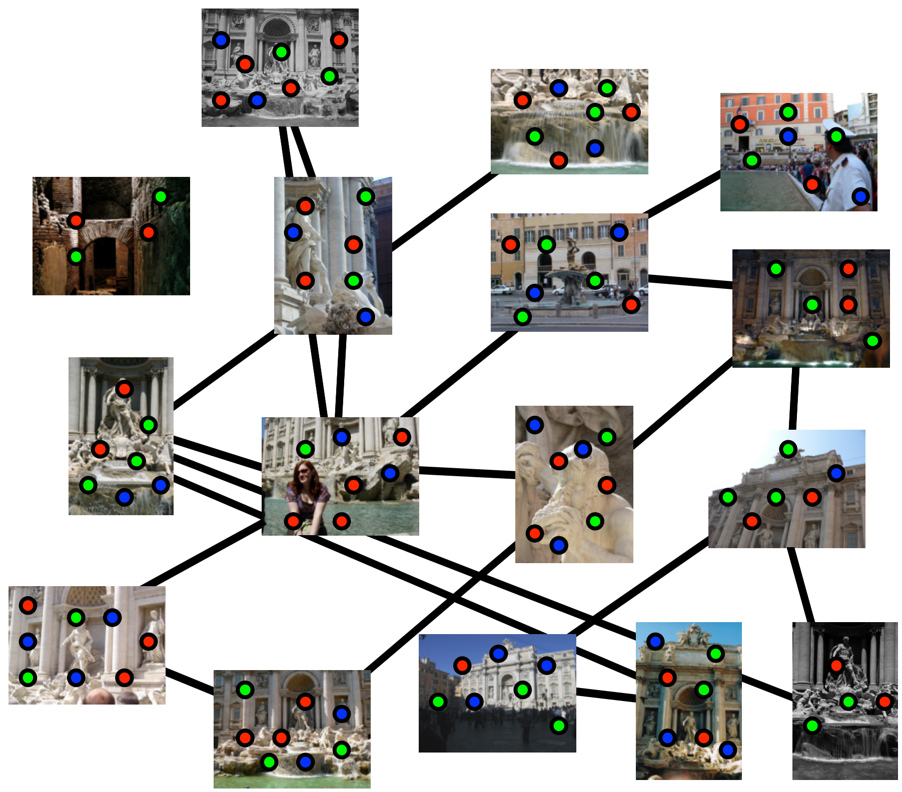
Feature matching

Match features between each pair of images



Feature matching

Refine matching using RANSAC to estimate fundamental matrix between each pair



Correspondence estimation

- Link up pairwise matches to form connected components of matches across several images: tracks

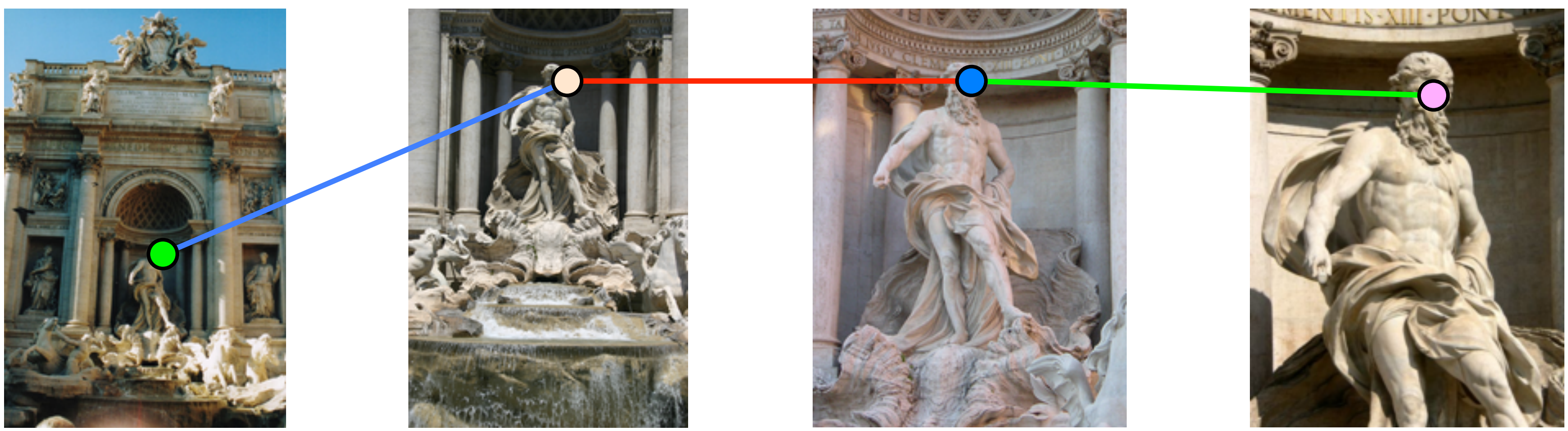
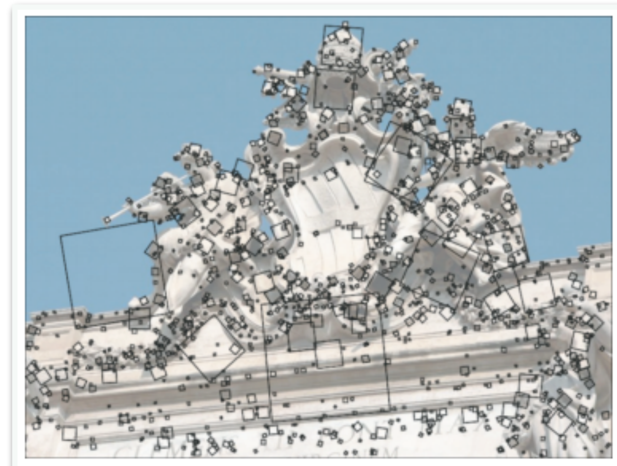


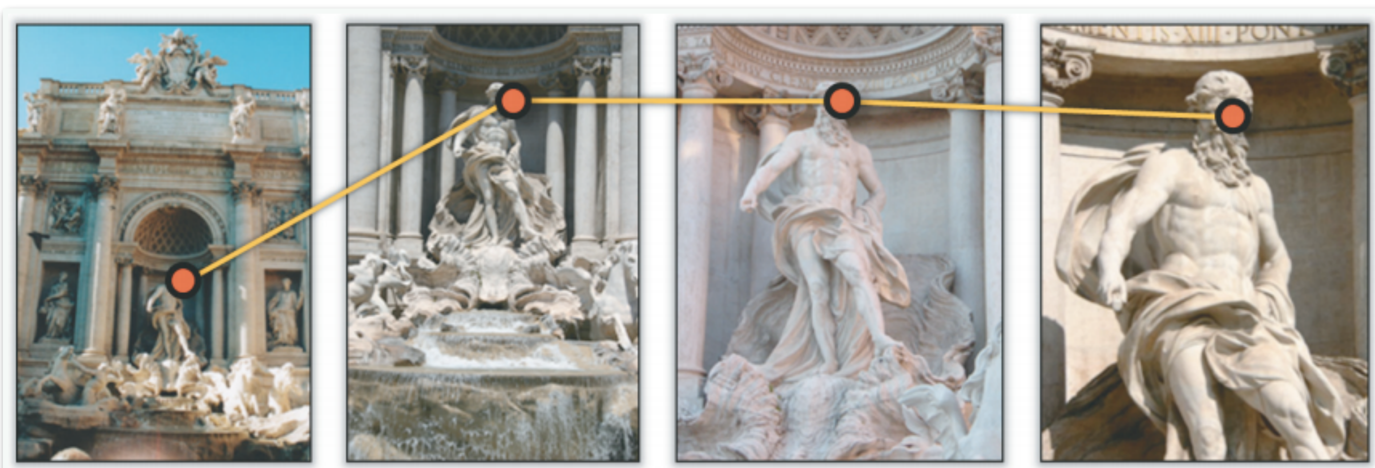
Image 1 Image 2 Image 3 Image 4

Geometric inference based on tracks

Photos

Features [1]





Tracks [2]

[1] David Lowe, "Distinctive image features from scale-invariant keypoints". IJCV 2004.

[2] N. Snavely, S. Seitz, R. Szeliski, "Photo tourism: exploring photo collections in 3D". SIGGRAPH 2006.

Structure from motion

- Minimize sum of squared reprojection errors:

$$g(\mathbf{X}, \mathbf{R}, \mathbf{T}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \underbrace{w_{ij}}_{\text{is point } i \text{ visible in image } j\text{ ?}} \cdot \left\| \underbrace{\mathbf{P}(\mathbf{x}_i, \mathbf{R}_j, \mathbf{t}_j)}_{\substack{\textit{predicted} \\ \text{image location}}} - \underbrace{\begin{bmatrix} u_{i,j} \\ v_{i,j} \end{bmatrix}}_{\substack{\textit{observed} \\ \text{image location}}} \right\|^2$$

- Minimizing this function is called *bundle adjustment*
  - Optimized using non-linear least squares, e.g. Levenberg-Marquardt

Problem size

Trevi Fountain collection

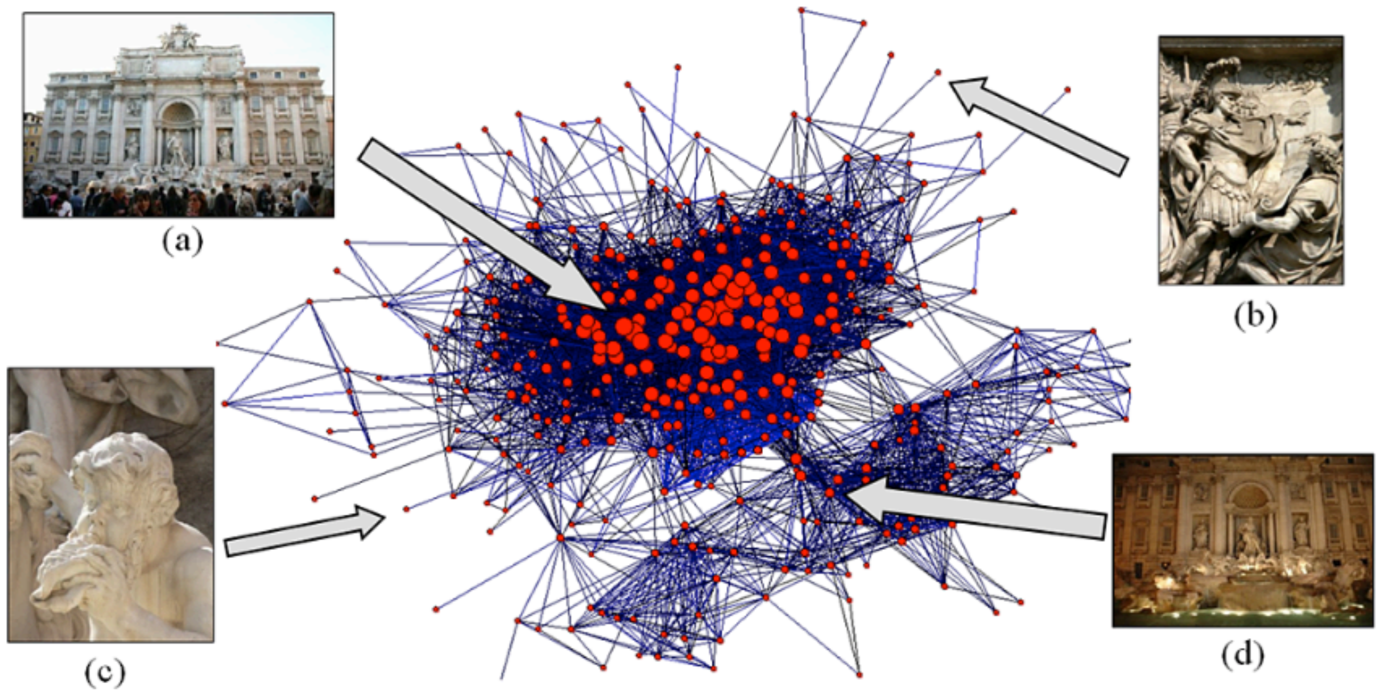
466 input photos

+ > 100,000 3D points

= very large optimization problem

Image connectivity graph

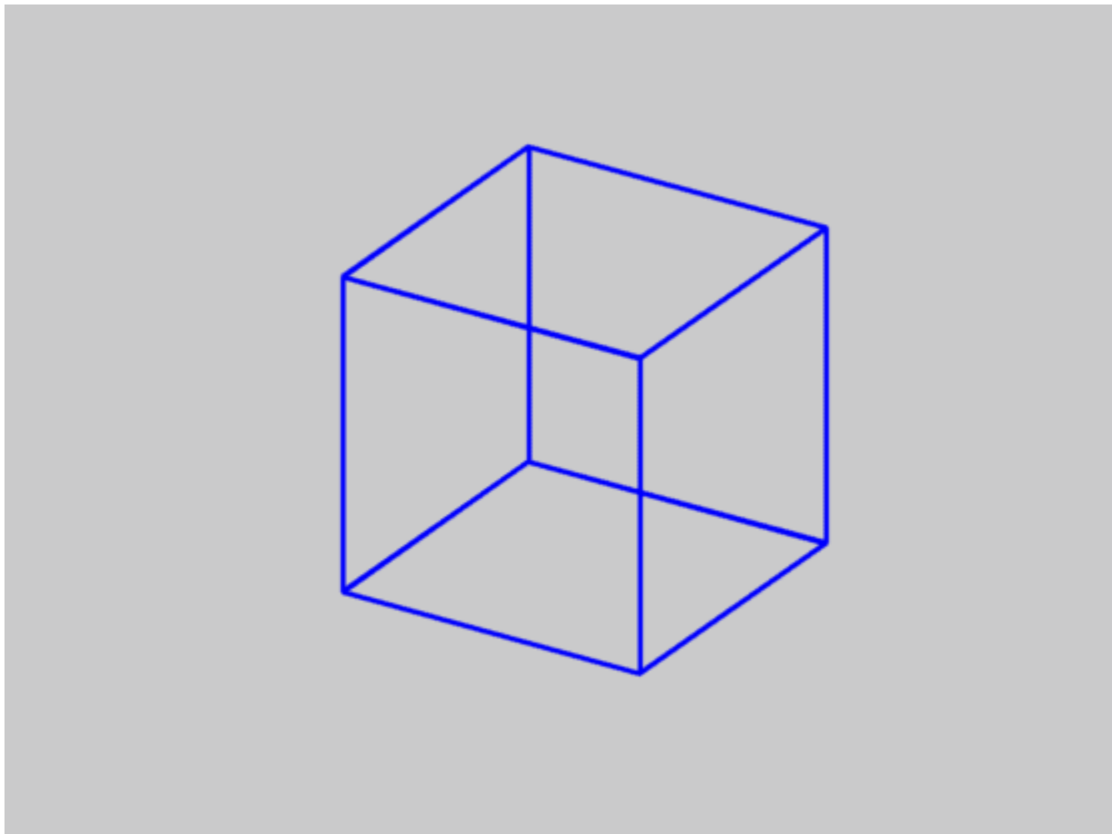
Fig. 1 Photo connectivity graph. This graph contains a node for each image in a set of photos of the Trevi Fountain, with an edge between each pair of photos with matching features. The size of a node is proportional to its degree. There are two dominant clusters corresponding to day (**a**) and night time (**d**) photos. Similar views of the facade cluster together in the center, while nodes in the periphery, e.g., (**b**) and (**c**), are more unusual (often close-up) views



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

Is SfM always uniquely solvable?

- No…



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 in the image remain exactly the same:

$$\mathbf{x} = \mathbf{P}\mathbf{X} = \left(\frac{1}{k}\mathbf{P}\right)(k\,\mathbf{X})$$

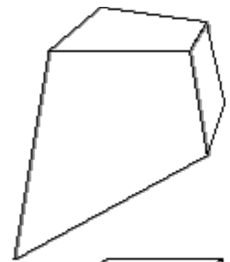
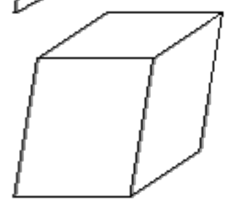
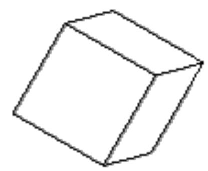
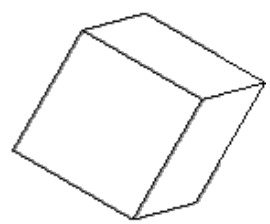
It is impossible to recover the absolute scale of the scene!

Structure from motion ambiguity

- More generally: if we transform the scene using a transformation **Q** and apply the inverse transformation to the camera matrices, then the images do not change

$$\mathbf{x} = \mathbf{P}\mathbf{X} = \left(\mathbf{P}\mathbf{Q}^{-1}\right)\left(\mathbf{Q}\mathbf{X}\right)$$

Types of ambiguity

Projective 15dof	$\begin{bmatrix} A & t \\ v^T & v \end{bmatrix}$		Preserves intersection and tangency
Affine 12dof	$\begin{bmatrix} A & t \\ 0^T & 1 \end{bmatrix}$		Preserves parallellism, volume ratios
Similarity 7dof	$\begin{bmatrix} sR & t \\ 0^T & 1 \end{bmatrix}$		Preserves angles, ratios of length
Euclidean 6dof	$\begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}$		Preserves angles, lengths

Structure from Motion

- Repetitive structures

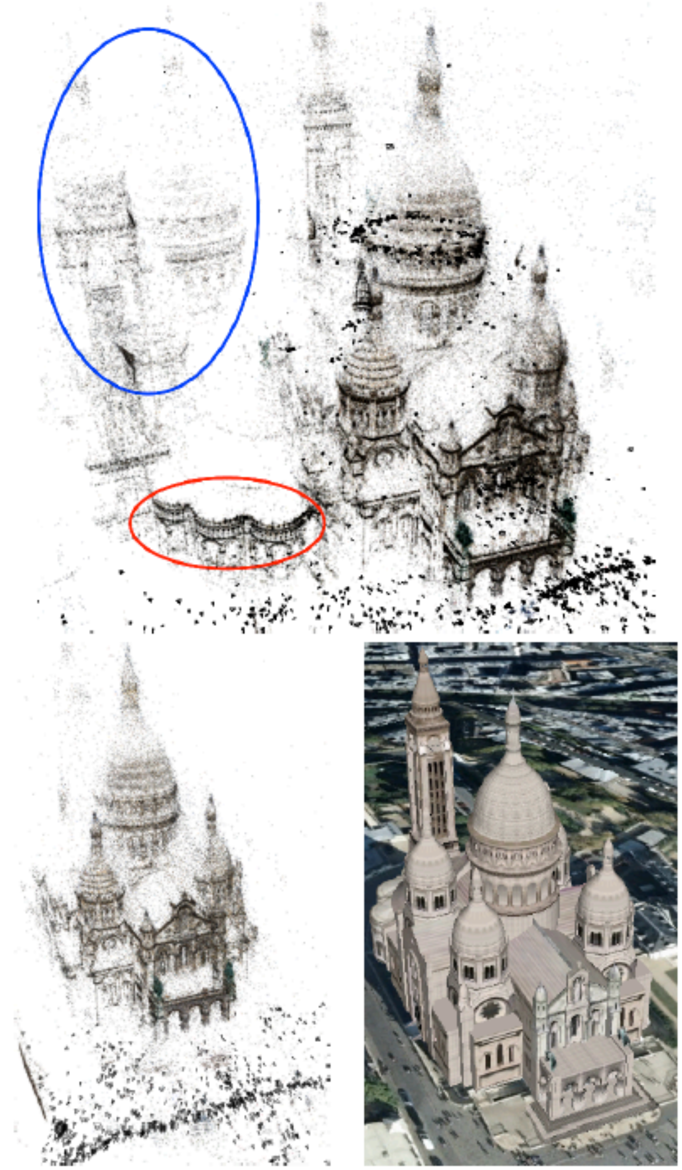
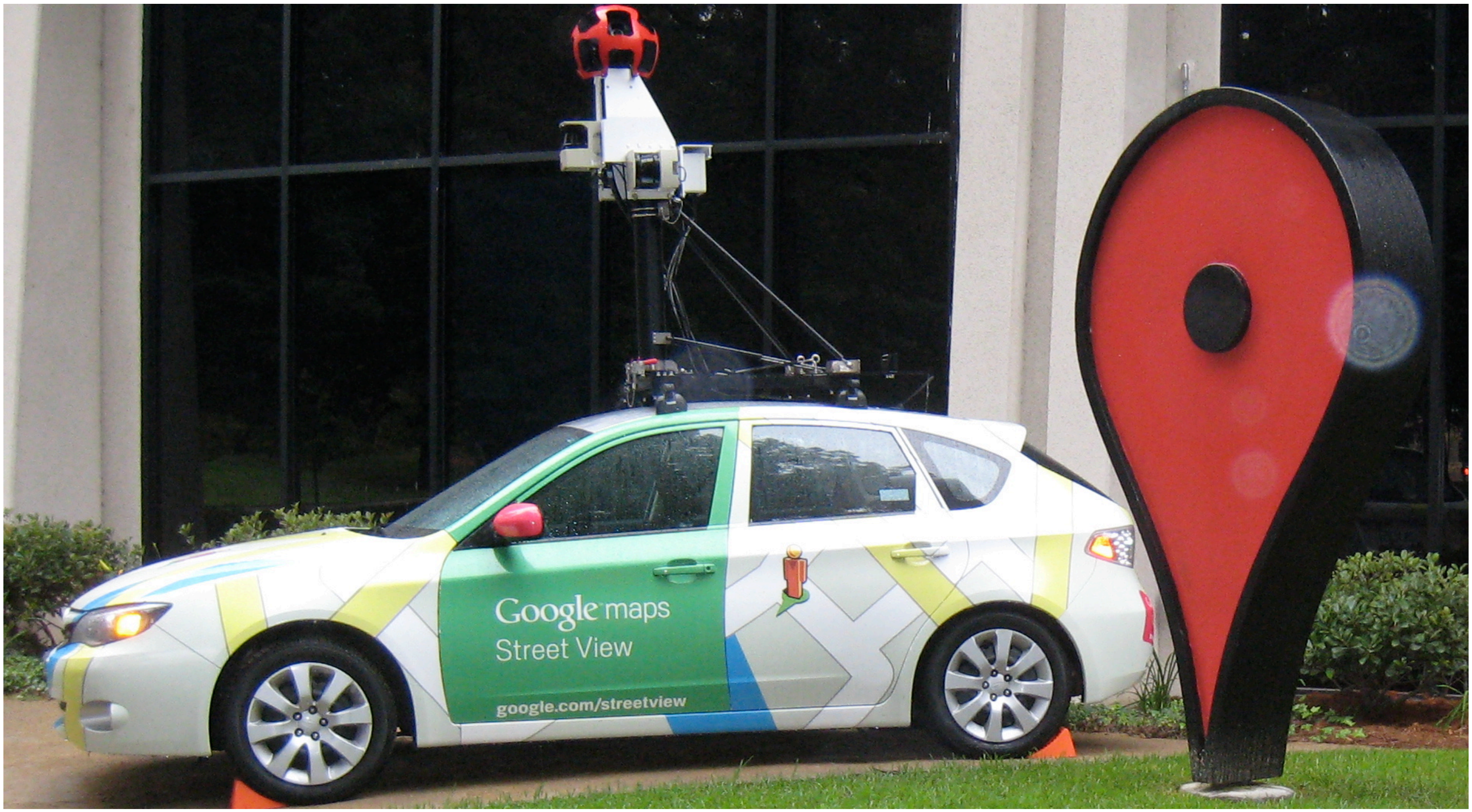


Figure 1. (a) A SfM model of the Sacre Coeur Basilica in Paris containing structural ambiguities. Prominent errors in the reconstruction, including repeated and phantom structures, are highlighted. (b) The same model, correctly disambiguated using our proposed method. (c) A Google Earth rendering of the cathedral.

SfM applications

- 3D modeling
- Surveying
- Robot navigation and mapmaking
- Archeological reconstruction
- Visual effects

Google Street View



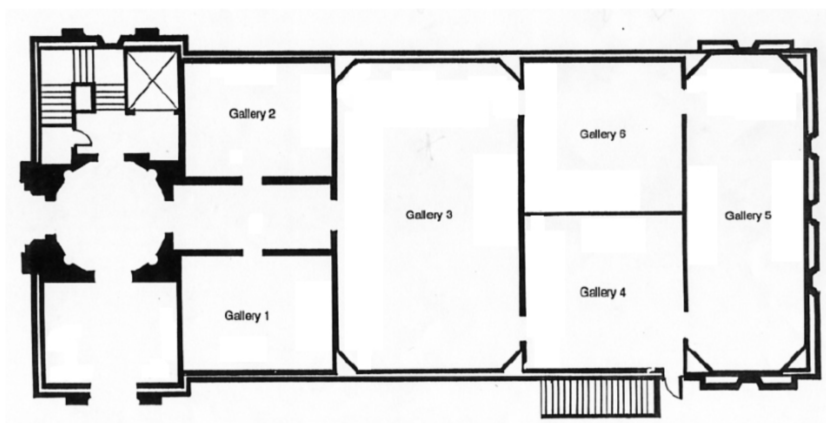
Visual Turing Test

Figure 5. Visual Turing test. In each image pair, the ground truth image is on the left and our result is on the right.

Challenges – Indoor Reconstruction

Texture-poor surfaces

Complicated visibility



Prevalence of thin structures
(doors, walls, tables)

Museums

Google Art

State-of-the-art

- sFM used for large-scale internet level 3D reconstruction

- Future: expect 3D imagery
  - Kinect, sFM, etc.

Where are we?

- First: low-level vision, features
- Second: 3D reconstruction
- Next: Recognition
- Last few lectures: Computational photography