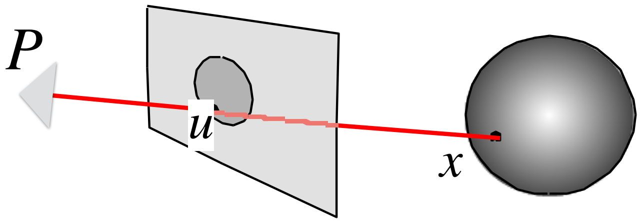
3D Scene Reconstruction From Images

Luiz Velho
IMPA

Mathematical Fundamentals

- Projective Geometry (Recap)

$$2D \longleftrightarrow 3D$$



$$u = Px \qquad\qquad u \in \mathbb{R}^2, \quad x \in \mathbb{R}^3$$

Vision Problems

$$u_i = P x_i$$

$$u = P x$$

camera

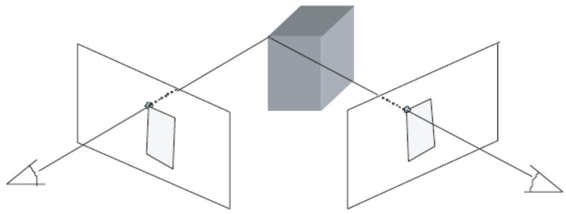
- $\boxed{\text{Reconstruction}}$ $u = P x$

3D scene

Multiview 3D Reconstruction

Multiview 3D Reconstruction

> *Several Images - One Scene*



Calibration

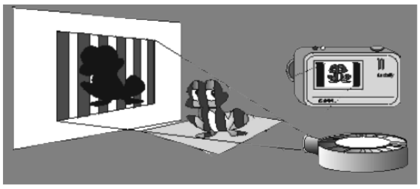
- Approach
  - View Based

Main Approaches

- View-Based Reconstruction
  - Graphics (Known Cameras)
- Structure from Motion
  - Vision (Unknown Cameras)

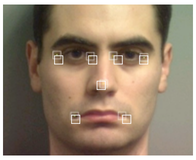
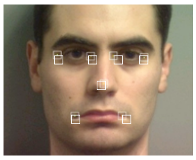
View-Based Reconstruction

- Classification
  - Active
    (Structured Light)



Correspondence

  - Passive (SFM)
    (Multi-View Stereo)

Structured Light Reconstruction

Active Stereo

OBS: <u>Known</u> Cameras

Structured-Light Reconstruction

Processing Pipeline (Operations)

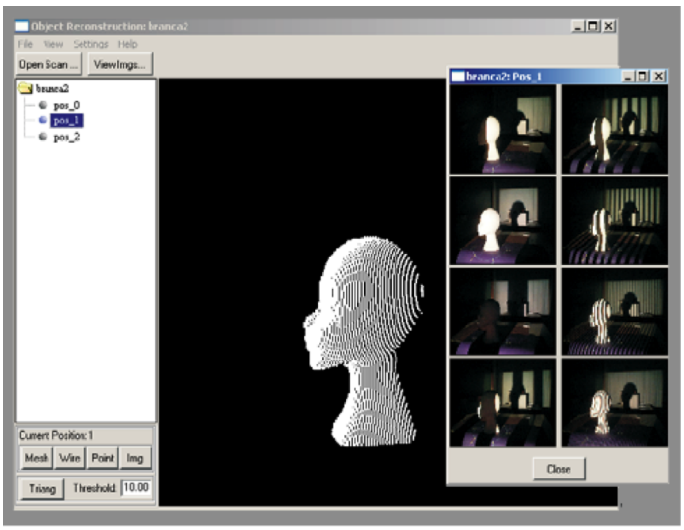
Structured Light → Triangulate
...
Structured Light → Triangulate

→ Align → Merge → Structure / Appearance

3D Scene Reconstruction

Correspondences · *3D Points* · *meshes* · *volume* · *mesh + Texture*

Structured Light



stripe boundary code

Triangulation

- Given Cameras, P_k and Assuming Correspondences, $u_i^{(k)}$

factor $P_k = K_k[R_k \mid \mathbf{t}_k]$,

compute the camera center and a world-direction ray for pixel $\mathbf{u}_i^{(k)}$

$$\mathbf{C}_k = -R_k^\top \mathbf{t}_k, \qquad \mathbf{d}_k = \frac{R_k^\top K_k^{-1} \mathbf{u}_i^{(k)}}{\| R_k^\top K_k^{-1} \mathbf{u}_i^{(k)} \|}.$$

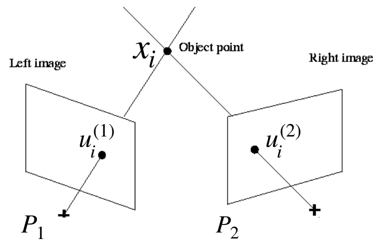
Find scalars λ_1, λ_2 that minimize $\|\mathbf{C}_1 + \lambda_1 \mathbf{d}_1 - (\mathbf{C}_2 + \lambda_2 \mathbf{d}_2)\|^2$.

Let $\mathbf{w}_0 = \mathbf{C}_1 - \mathbf{C}_2$, $a = \mathbf{d}_1 \cdot \mathbf{d}_1$, $b = \mathbf{d}_1 \cdot \mathbf{d}_2$, $c = \mathbf{d}_2 \cdot \mathbf{d}_2$, $d = \mathbf{d}_1 \cdot \mathbf{w}_0$, $e = \mathbf{d}_2 \cdot \mathbf{w}_0$.

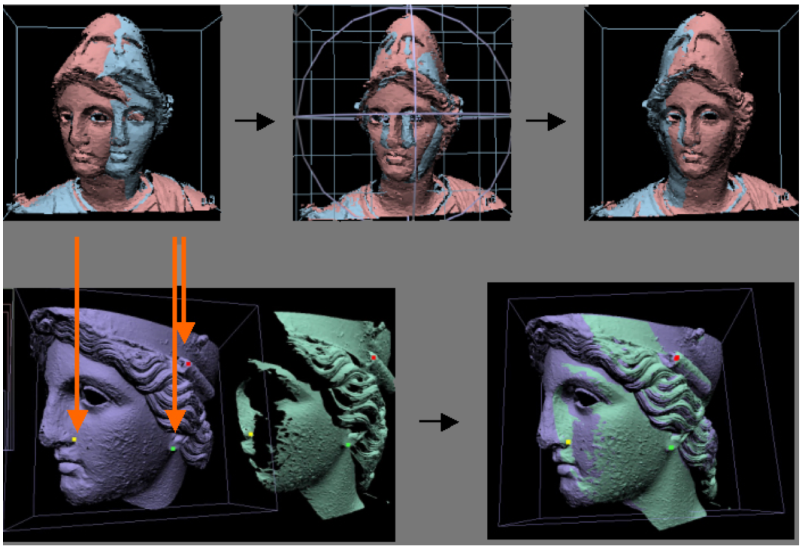
$$\lambda_1 = \frac{be - cd}{ac - b^2}, \qquad \lambda_2 = \frac{ae - bd}{ac - b^2},$$

and a common estimate is the midpoint

$$\mathbf{x}_i = \frac{1}{2}\left(\mathbf{C}_1 + \lambda_1 \mathbf{d}_1 + \mathbf{C}_2 + \lambda_2 \mathbf{d}_2\right).$$

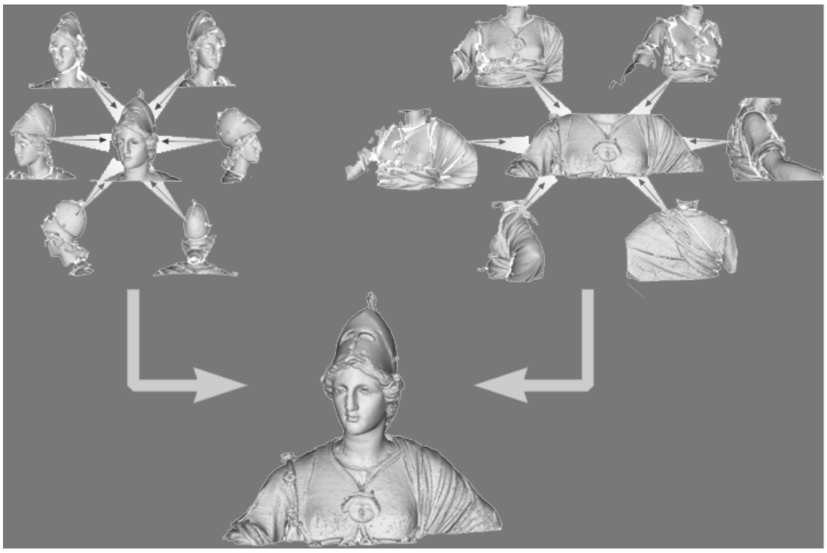


Mesh Align



Iterated Closest Point (ICP)

Mesh Merge



Truncated Signed Distance Function (TSDF)

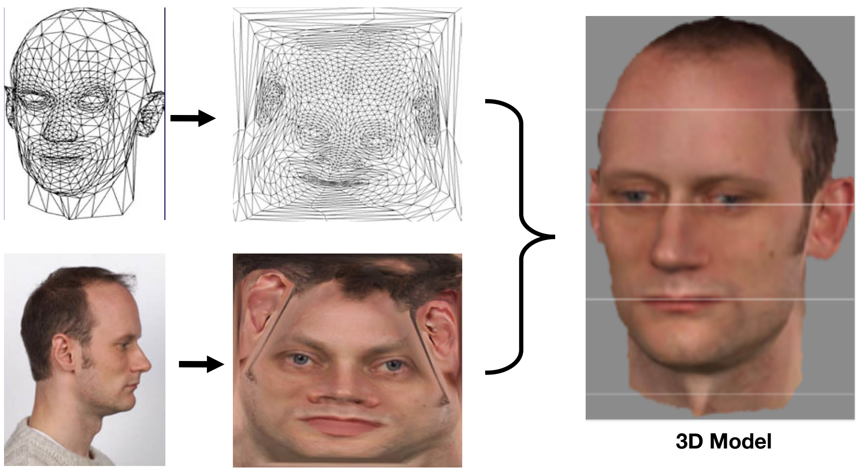
Appearance / Attribute Modeling

- (u,v) Mapping
  * Texture Atlas

- BRDF Estimation
  * *Unshading*

- Normal Mapping
  * Shape from Shading

Textured Model

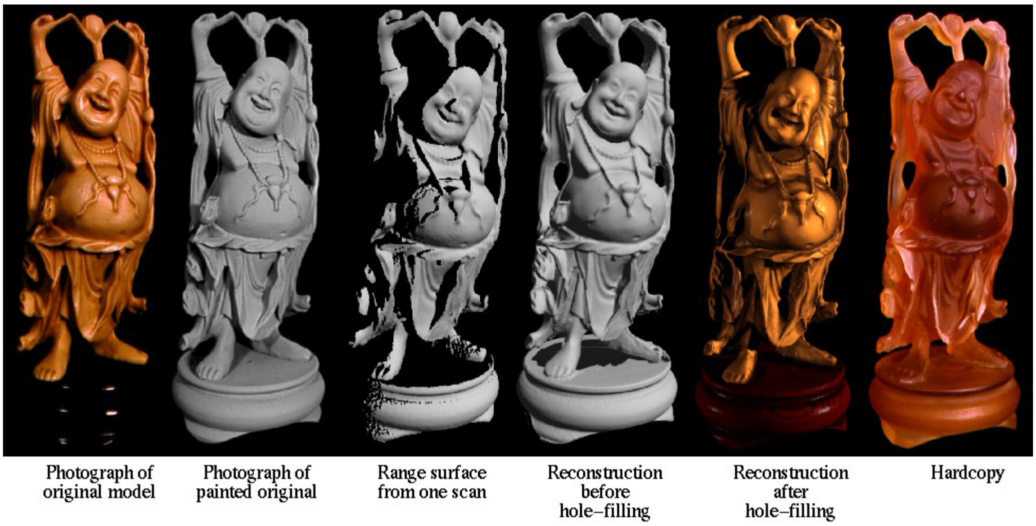


surface parametrization + mapping

3D Model

Surface Properties

Complete Pipeline



original **range** **raw mesh** **appearance** **processed**

(Mark Levoy)

Structure-from-Motion

Passive Multi-View Stereo

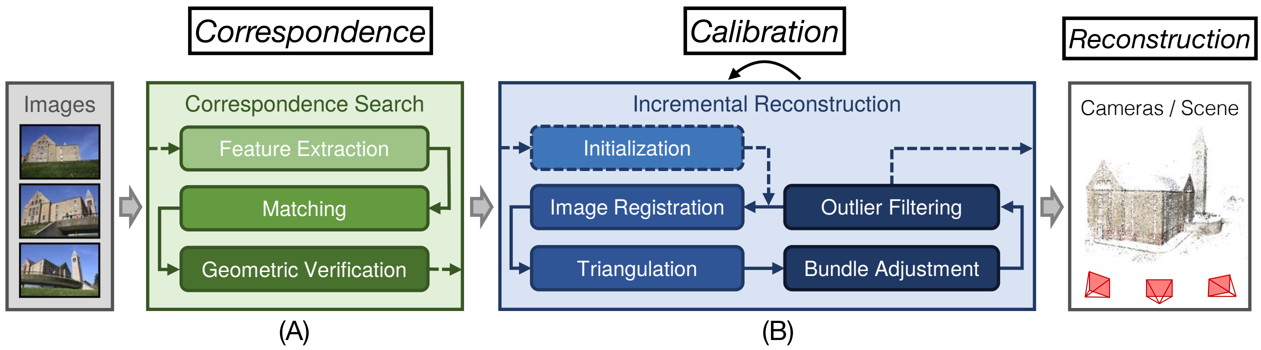
OBS: <u>Unknown</u> Cameras

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

COLMAP

Incremental Structure from Motion

Incremental Structure-from-Motion



Feature matching and reconstruction pipeline (COLMAP):

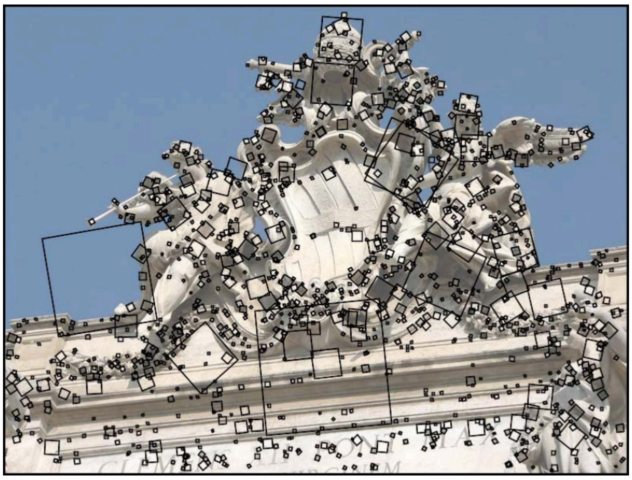
(A) - **Correspondence Search:** Find and match robust 2D features acroess images

(B) - **Incremental Reconstruction:** Start with 2 views, incrementally add cameras

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Feature Extraction

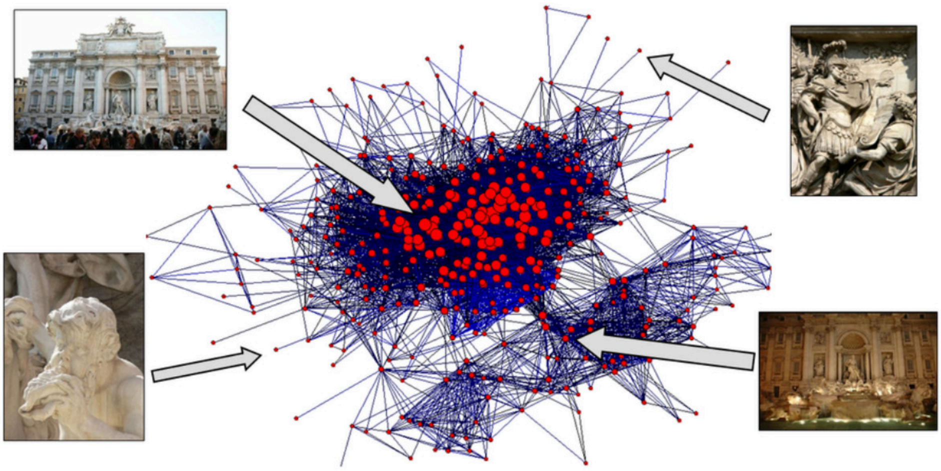


► **Detect features** (eg, SIFT, SURF, BRISK) in all input images

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Feature Matching & Geometric Verification



► Find **overlapping image pairs** and **associated feature correspondences** (i.e., graph) (RANSAC)

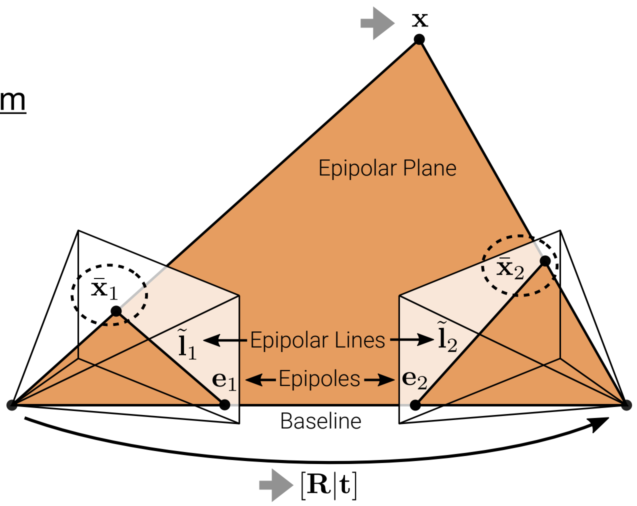
Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Initialization

- Eight-Point Algorithm

$$\tilde{\mathbf{x}}_2^T \mathbf{F} \tilde{\mathbf{x}}_1 = 0$$
$$\mathbf{F} = \mathbf{K}_2^{-T} (\mathbf{R}[\mathbf{t}]_\times) \mathbf{K}_1^{-1}$$



- ▶ Select **two views** with many correspondences and **estimate their geometry** $[R \mid t]$
  Leverage epipolar constraint to estimate $\mathbf{F}$ *from correspondences!* $(\tilde{x}_1, \tilde{x}_2)$

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Image Registration

Find a **new image** with correspondences to the current set and **estimate camera pose:**

Let $\mathcal{X} = \{\bar{\mathbf{x}}_i^s, \bar{\mathbf{x}}_i^w\}_{i=1}^N$ be a set of N 3D-to-2D correspondences related by $\bar{\mathbf{x}}_i^s = \mathbf{P}\,\bar{\mathbf{x}}_i^w$.

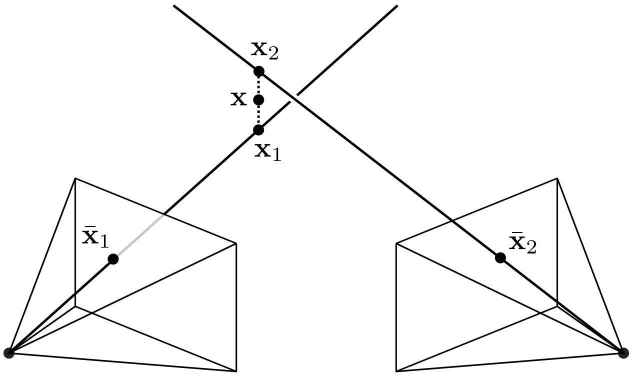
Thus, the equation above can be expressed as $\bar{\mathbf{x}}_i^s \times \mathbf{P}\bar{\mathbf{x}}_i^w = \mathbf{0}$.
Using the **Direct Linear Transform,** this can be written as a linear equation in the entries of $\mathbf{P}$. The solution to the constrained system (fixing $\mathbf{P}$'s scale) is given by SVD.

- ▶ Given that $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ and $\mathbf{K}$ is upper-triangular, both $\mathbf{K}$ and $\mathbf{R}$ can be easily obtained from the front 3×3 submatrix of $\mathbf{P}$ using standard RQ factorization
- ▶ If $\mathbf{K}$ is known, we can even estimate $\mathbf{P}$ from only three points (P3P algorithm)
- ▶ In practice, random sampling consensus (RANSAC) is used to remove outliers

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted frc Adapted from: University of Tübingen: Computer Vision, Prof. Andreas G

Triangulation



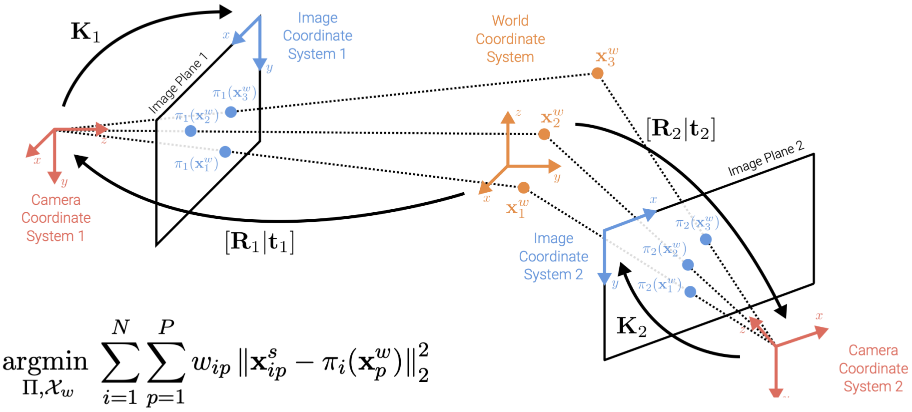
- Given the newly registered image, **new correspondences** can be **triangulated**
- In COLMAP, a robust triangulation method is proposed that handles outliers

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Bundle Adjustment & Outlier Filtering

Minimize reprojection error of all observations wrt. all cameras and 3D points: Π^*, X_w^*



$$\Pi^*, \mathcal{X}_w^* = \underset{\Pi, \mathcal{X}_w}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{p=1}^{P} w_{ip} \left\| \mathbf{x}_{ip}^s - \pi_i(\mathbf{x}_p^w) \right\|_2^2$$

$\mathbf{K}_i$ and $[\mathbf{R}_i|\mathbf{t}_i]$ are the intrinsic and extrinsic parameters of π_i, respectively.
During bundle adjustment, we optimize $\{(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)\}$ and $\{\mathbf{x}_p^w\}$ jointly.

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

Bundle Adjustment & Outlier Filtering (Summary)

Minimize reprojection error of all observations wrt. all cameras and 3D points:

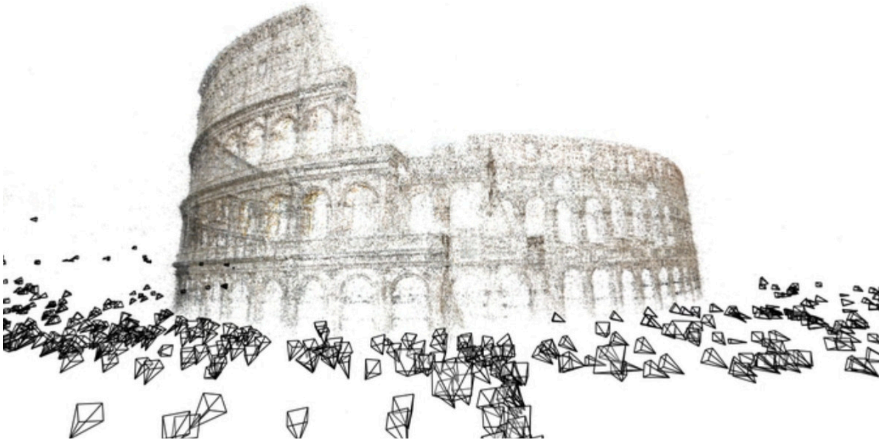
$$\Pi^*, \mathcal{X}_w^* = \underset{\Pi, \mathcal{X}_w}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{p=1}^{P} w_{ip} \left\| \mathbf{x}_{ip}^s - \pi_i(\mathbf{x}_p^w) \right\|_2^2$$

- Since incremental SfM only affects the model locally, COLMAP performs **local BA** on the locally connected images, and **global BA** only once in a while for efficiency
- For solving the sparse large-scale optimization problem, COLMAP uses Ceres
- After each BA, observations with large reprojection errors and cameras with abnormal field of views or large distortion coefficients are removed

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

Results and Applications

Building Rome in a Day



- First large-scale SfM from unstructured images: **Building Rome in a Day**

Agarwal, Snavely, Simon, Seitz and Szeliski: Building Rome in a day. ICCV, 2009.

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

COLMAP SfM

- **COLMAP** significantly improves accuracy and robustness compared to prior work

Schönberger and Frahm: Structure-from-Motion Revisited. CVPR, 2016.

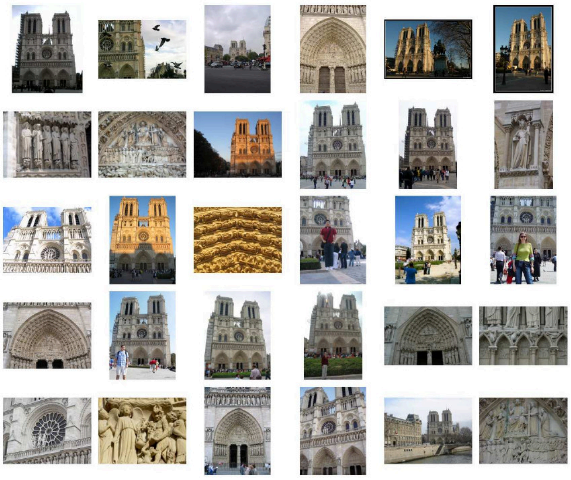
Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

COLMAP MVS

- COLMAP features a second **multi-view stereo stage** to obtain dense geometry (see Structured Light Pipeline)

Schönberger, Zheng, Frahm and Marc Pollefeys: Pixelwise View Selection for Unstructured Multi-View Stereo. ECCV, 2016.

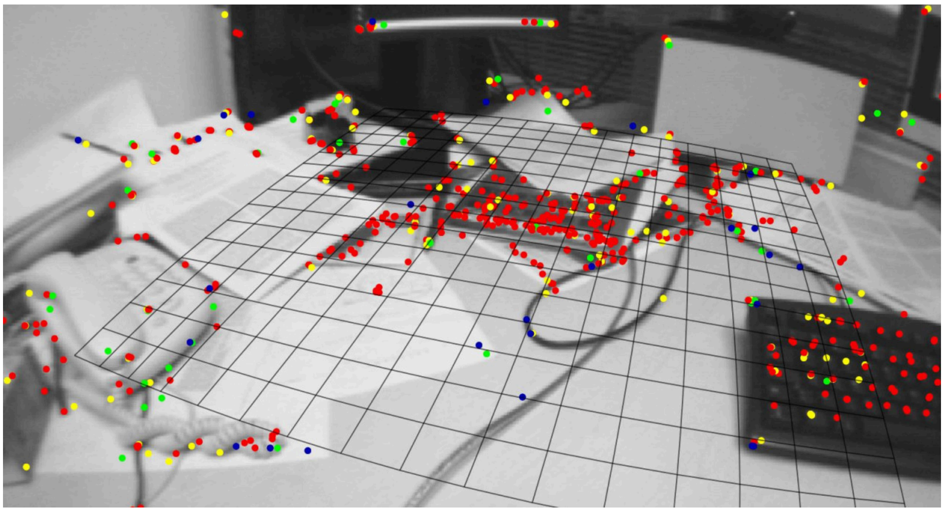
Photo Tourism



- **Photo Tourism / PhotoSynth** allows for exploring photo collections in 3D (*Multiview Interpolation for 3D Points with Color*)

Snavely, Seitz and Szeliski: Photo tourism: exploring photo collections in 3D. SIGGRAPH, 2006.

Parallel Tracking and Mapping (PTAM)



► **PTAM** demonstrates real-time tracking and mapping of small workspaces

Klein and Murray: Parallel Tracking and Mapping for Small AR Workspaces. ISMAR, 2007.