Automatic key frame selection for video-based structure from motion pipelines

Mauriana Pesaresi Seminar series 2022

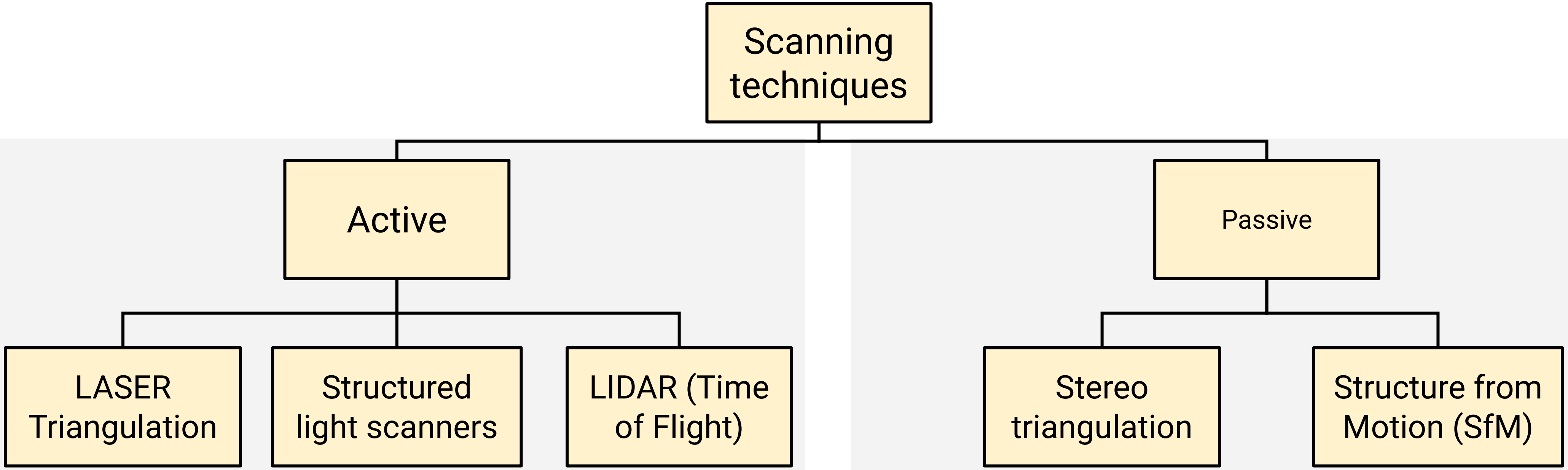
ARSLAN SIDDIQUE

Marie Curie PhD Fellow, Visual Computing lab, ISTI-CNR

Email : a.siddique@phd.unipi.it

Supervisor: Paolo Cignoni

3D Image acquisition



Scanning techniques

Active

LASER Triangulation

Structured light scanners

LIDAR (Time of Flight)

Passive

Stereo triangulation

Structure from Motion (SfM)

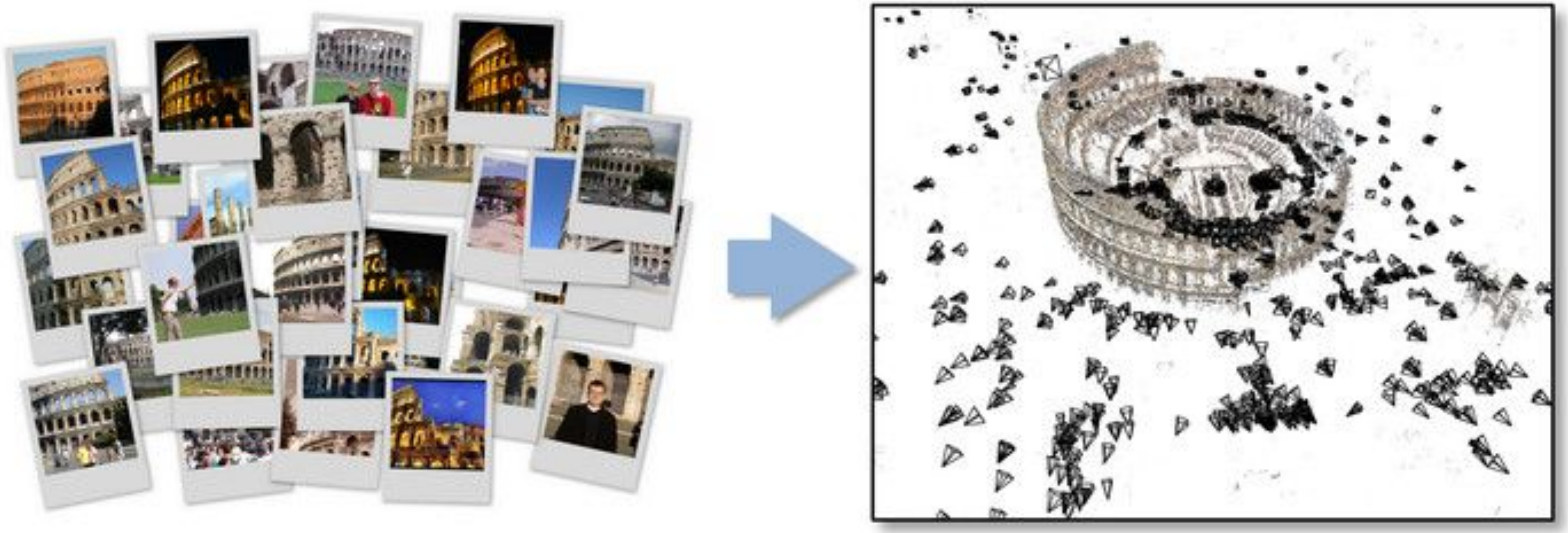
- Advantage is the high depth accuracy.
- Disadvantage is the cost and lack of color information in some cases)

- Advantage is the high amount of color information and low cost.
- Disadvantage is lesser accuracy of depth estimations.

- Multiple modalities (Active and passive) are combined to build accurate systems.

Structure from Motion (SfM)

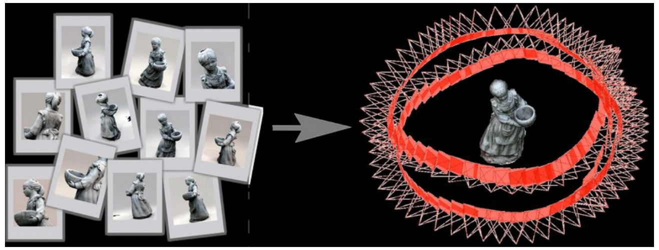
- Create a 3D object representation using using images from different viewpoints of the object.



Source: Photo tourism

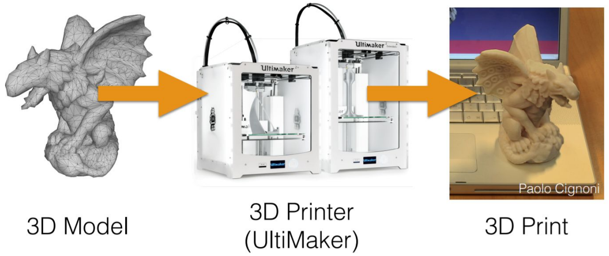
Applications

- Cultural heritage and Virtual museums



- Augmented reality and Metaverse

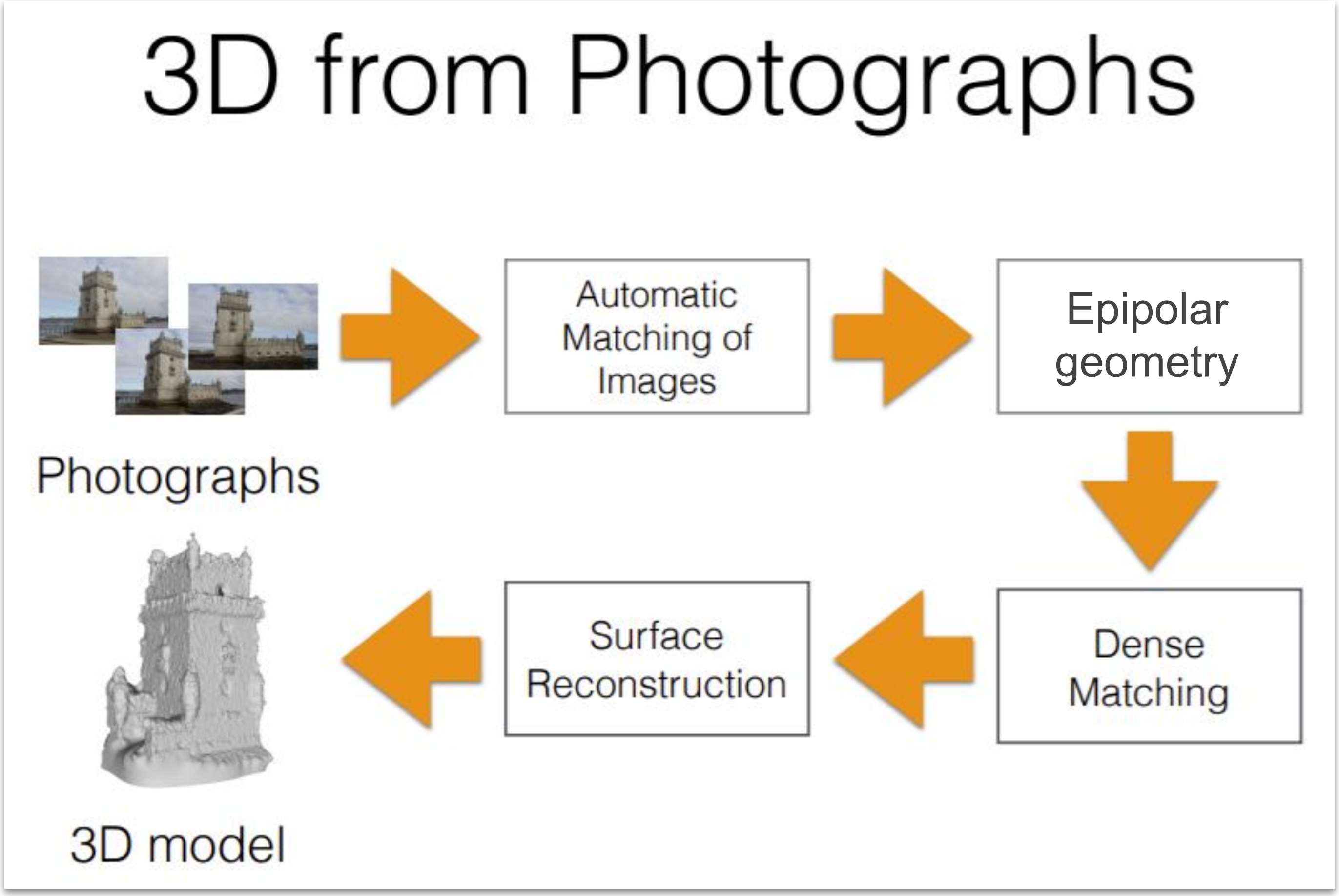
- 3D Printing



- Autonomous navigation and guidance

Structure from Motion
Initial 3D map generation

SfM steps



3D from Photographs

Photographs

Automatic Matching of Images

Epipolar geometry

Dense Matching

Surface Reconstruction

3D model

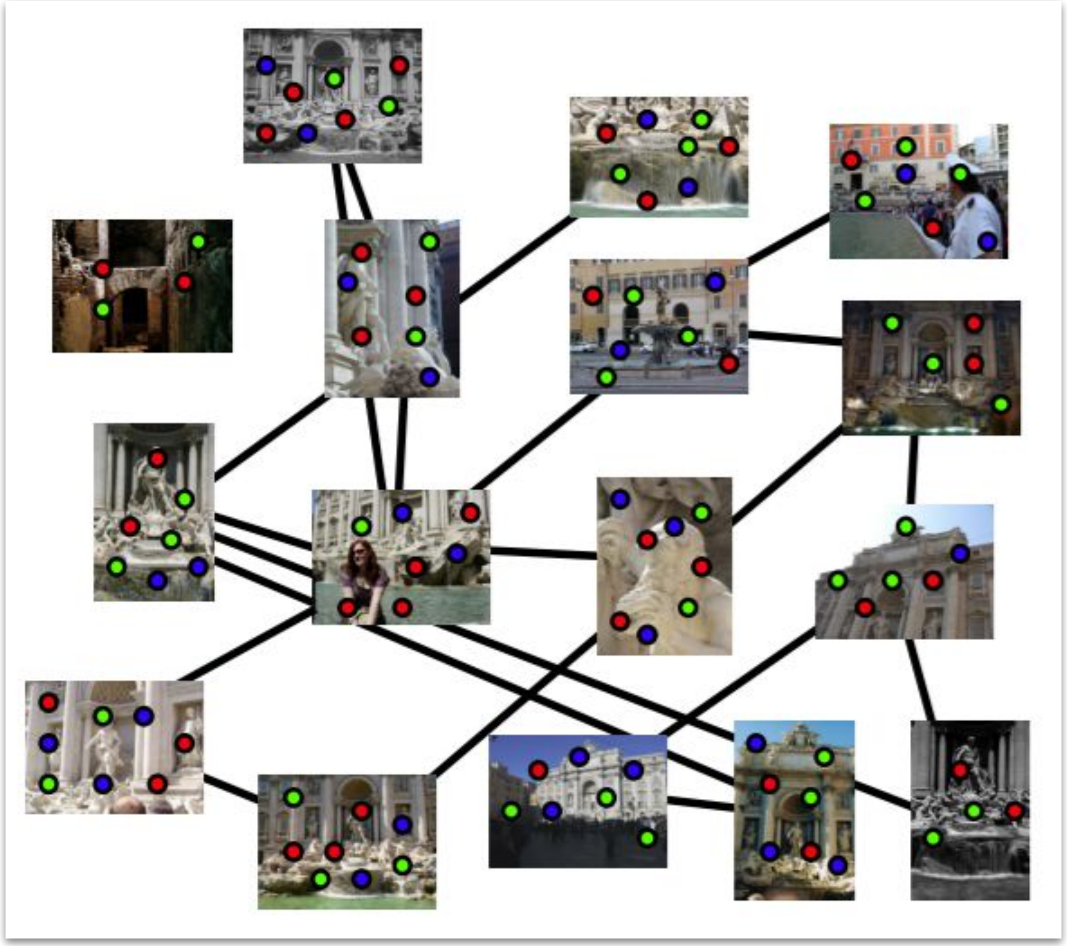
Source: Course

SfM Steps

- Feature Matching.

- Examples include SIFT/SURF features, Harris Corners etc.

- OpenCV has built in methods to compute and match features.



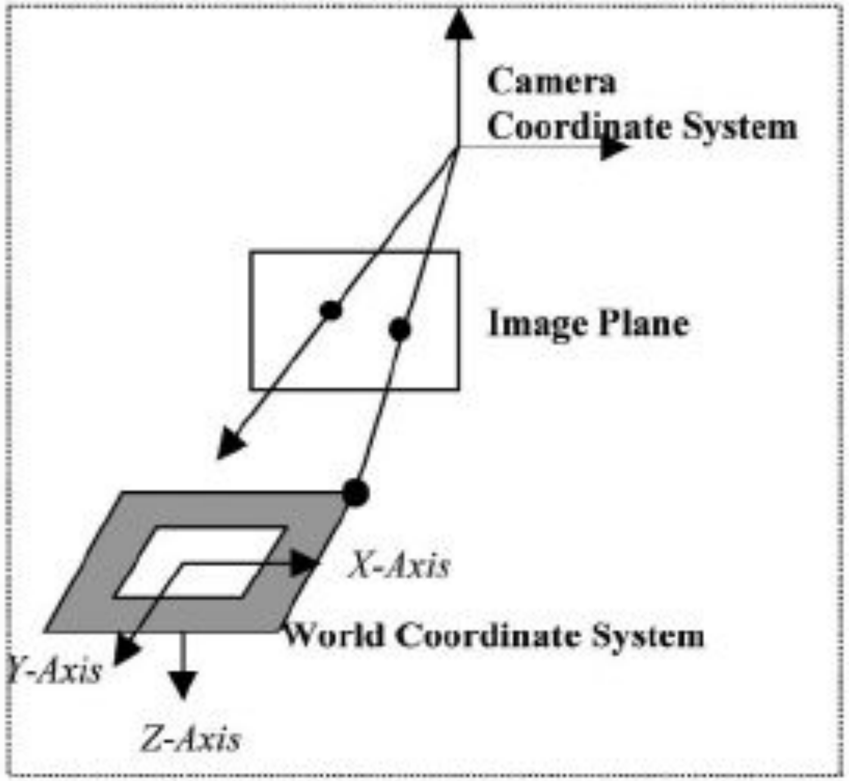
Source : Slides

SfM Steps

- Camera calibration, 3D camera poses and 3D scene recovery.

- We should already know the intrinsic matrices of cameras.

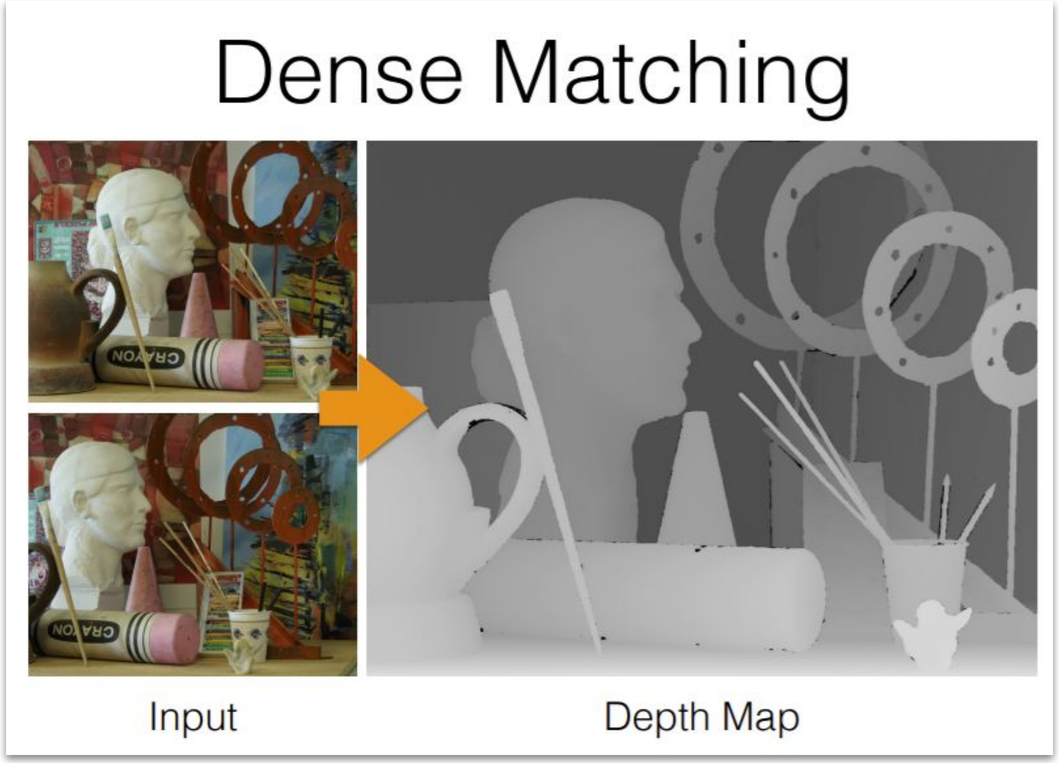
- Transformation from world coordinate system to camera coordinate system is called camera extrinsics or camera pose.



Source : Yuan *et. al.*

SfM Steps

- Depth map contains depth of every pixel and confidence level.



Dense Matching

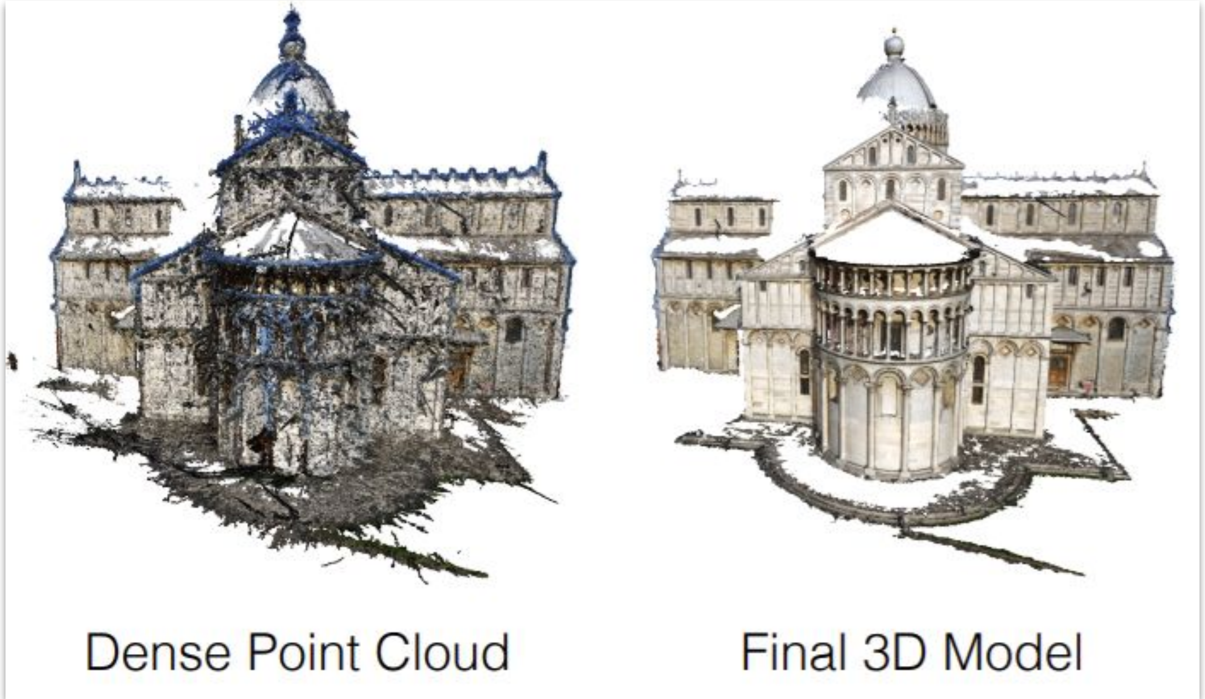
Input

Depth Map

Source: Course

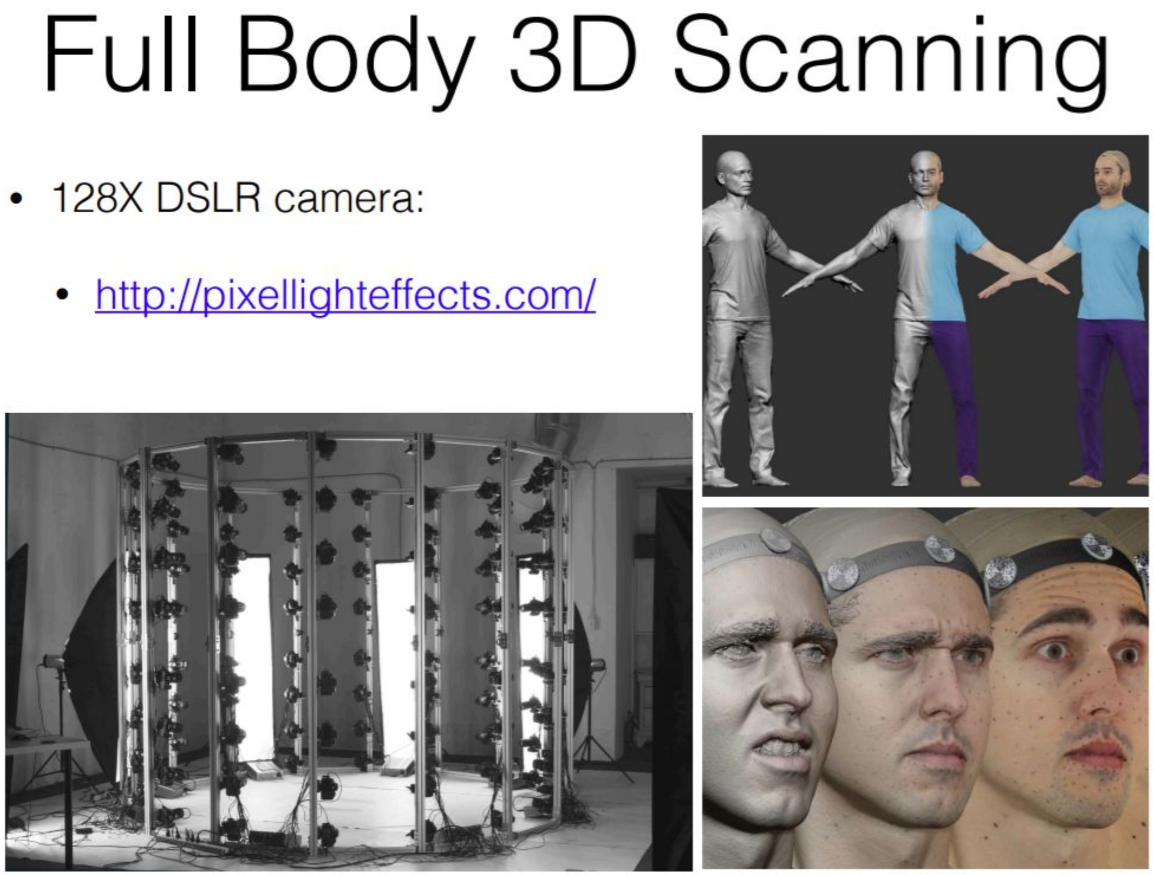
SfM Steps

- Combine depth maps to build dense point cloud.
- Surface reconstruction to build final model



Source: Course

Example



Full Body 3D Scanning

- 128X DSLR camera:
  - http://pixellighteffects.com/

Source: Course

Softwares

- 51 SfM softwares (https://opensourcelibs.com/libs/structure-from-motion)
- COLMAP is very popular.
- It takes 2 hours and 40 minutes to process 1144 frames. A minute of video consists of 1800 frames (30*60).
- The quality of output model depends significantly on how you acquire photos.
- An aerial video of a rooftop gives poor results while rotating cameras around an object would give good result.

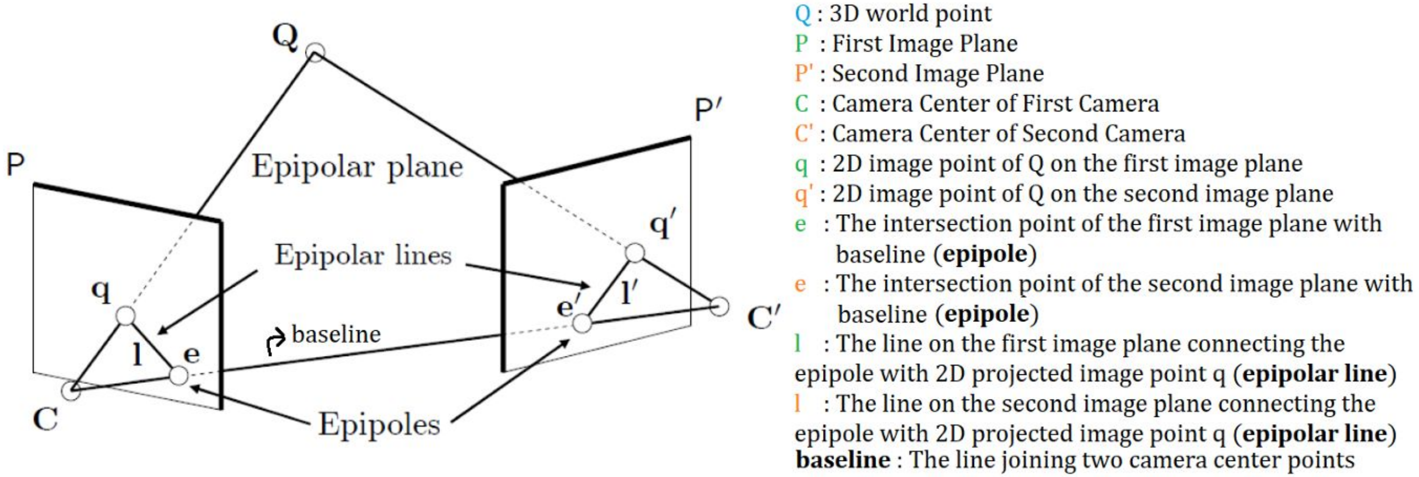
Video based SfM

- Video based SfM is challenging.
- Issues:
    a. Processing all frames is very expensive computationally.
    b. Inaccurate correspondences caused by motion blur
    c. Poor triangulation caused by small baseline in consecutive frames
    d. Degenerate cases (Fundamental matrix estimation fails if generality conditions do not hold)
- **Automatic selection of key-frames which give optimal reconstruction.**

Baseline and Triangulation

- Triangulation refers to the process of determining a point in 3D space give it’s projections on 2D images.
  a. If camera centers (C and C’) are very close to each other there would be no separate epipoles (e and e’) and Hence, coordinates of point in 3D will not be computed.

  b. Accuracy of Triangulation ∝ Length of baseline

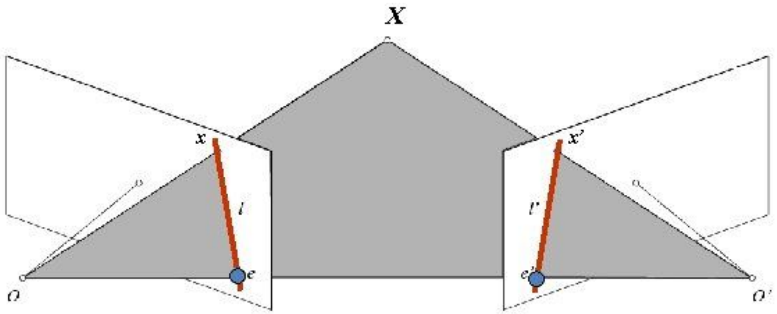


Q : 3D world point
P : First Image Plane
P' : Second Image Plane
C : Camera Center of First Camera
C' : Camera Center of Second Camera
q : 2D image point of Q on the first image plane
q' : 2D image point of Q on the second image plane
e : The intersection point of the first image plane with baseline (**epipole**)
e : The intersection point of the second image plane with baseline (**epipole**)
l : The line on the first image plane connecting the epipole with 2D projected image point q (**epipolar line**)
l : The line on the second image plane connecting the epipole with 2D projected image point q (**epipolar line**)
baseline : The line joining two camera center points

Source: Article

Fundamental Matrix

- Epipolar Constraint [1]
- F maps points in camera 1 to epipolar lines in camera 2.
- F connects the geometry of two cameras together.
- F is computed using 8 points algorithm.



$$x^T F x' = 0 \quad \text{with} \quad F = K^{-T} E K'^{-1}$$

$$\begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} x_1 x'_1 & x_1 y'_1 & x_1 & y_1 x'_1 & y_1 y'_1 & y_1 & x'_1 & y'_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_m x'_m & x_m y'_m & x_m & y_m x'_m & y_m y'_m & y_m & x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{21} \\ f_{31} \\ f_{12} \\ f_{22} \\ f_{32} \\ f_{13} \\ f_{23} \\ f_{33} \end{bmatrix} = 0$$

Degenerate cases

- Degenerate cases are the cases where 8-points algorithm fails to estimate Fundamental matrix.

  a. Motion Degeneracy
     If the camera rotates about it’s own center with no translation, epipolar geometry is not defined and hence algorithm fails.

  b. Structure Degeneracy
     If all points are coplanar, the fundamental matrix cannot be uniquely determined from image correspondences alone.

- These are degenerate camera motions

Homography matrix

- Homography transforms point in one image plane to the point in another image plane.

$$s\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

- Degenerate camera motion is more fittingly described by Homography while general camera motion is defined by Fundamental matrix.

GRIC

- Geometric Robust Information Criterion (GRIC) [2]

$$GRIC = \sum_i \rho(e_i^2)_i + \lambda_1 dn + \lambda_2 k,$$

where $\rho(e_i^2)$ is a robust function

$$\rho(e_i^2) = \min(\frac{e_i^2}{\sigma^2}, \lambda_3(r-d))$$

- Values are taken from [3]
- r = Dimension of the data. r = 4 for 2D correspondences between two frames.
- n = Total number of matched points
- ƛ1 = log(r)
- ƛ2 = log(rn)
- ƛ3 = Limits residual error = 2
- k = Number of degrees of freedom in the model. k = 7 for Fundamental matrix or k=8 for Homography
- d=Number of dimensions modeled. d=3 for Fundamental and d=2 for Homography.
- σ=Variance of tracker position. σ=2.

GRIC

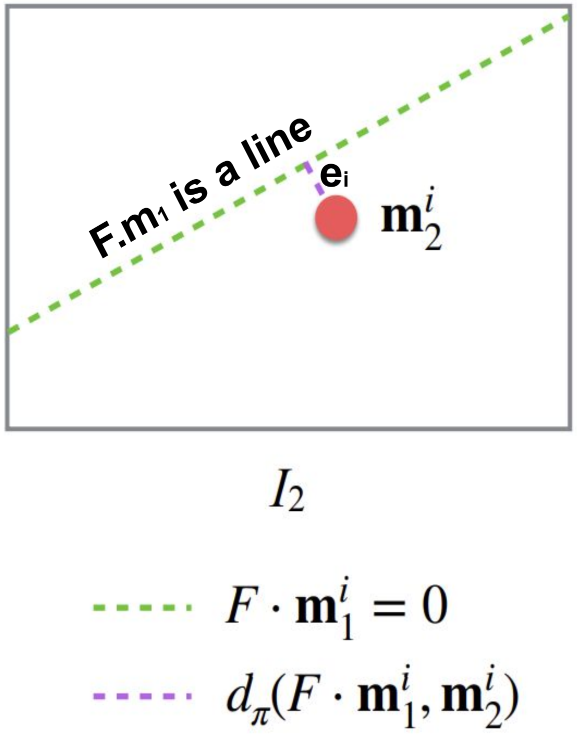
- Homography transforms point in 1 image plane to a point in another image plane. So, error is point to point distance.

  $e_{i,H} = x_2 - H^*x_1$

- Fundamental matrix transforms point in an image plane to a line in other image plane. So error is point to line distance.

  $e_{i,F}$ = point to line distance between F^*x_1 and x_2.

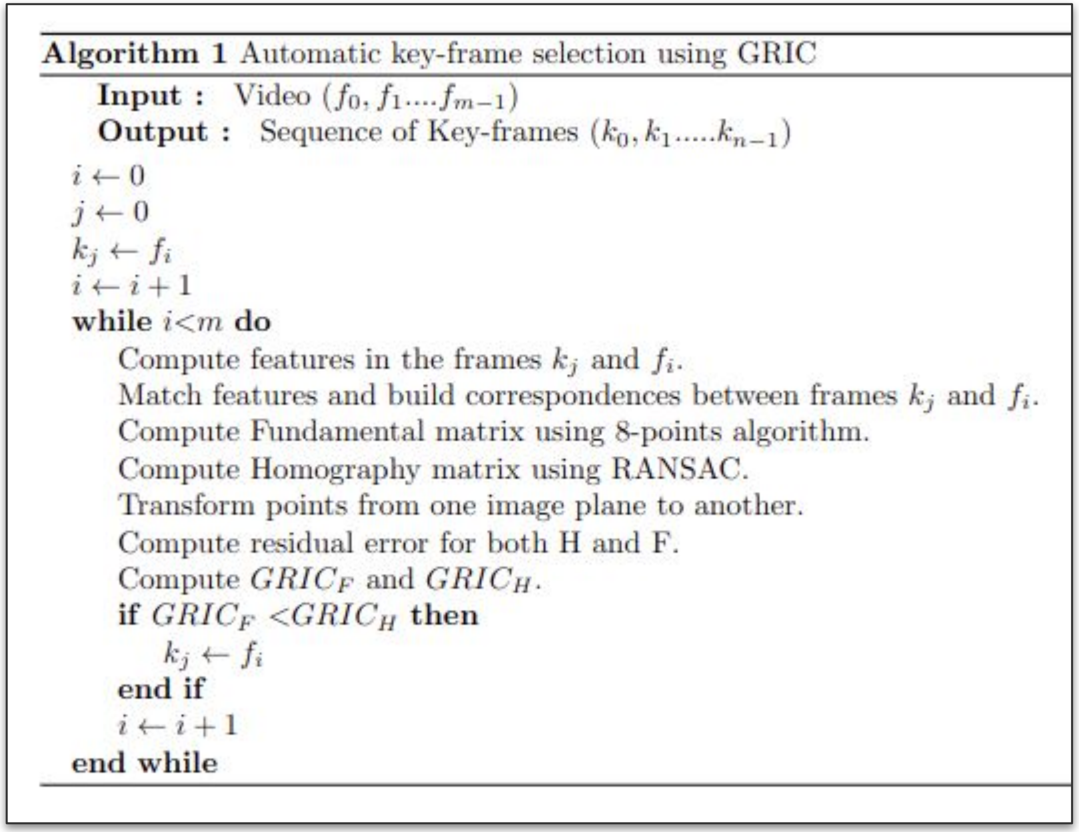
- Based upon these error values, $GRIC_H$ and $GRIC_F$ are computed.


F.m₁ is a line
eᵢ
$\mathbf{m}_2^i$
I_2
$F \cdot \mathbf{m}_1^i = 0$
$d_\pi(F \cdot \mathbf{m}_1^i, \mathbf{m}_2^i)$

Source: Course

Key-Frame selection algorithm [4]

- GRIC function is not available in OpenCV Python.



Algorithm 1 Automatic key-frame selection using GRIC

```
    Input :  Video ($f_0, f_1....f_{m-1}$)
    Output :  Sequence of Key-frames ($k_0, k_1.....k_{n-1}$)
  $i \leftarrow 0$
  $j \leftarrow 0$
  $k_j \leftarrow f_i$
  $i \leftarrow i + 1$
  while $i$<$m$ do
      Compute features in the frames $k_j$ and $f_i$.
      Match features and build correspondences between frames $k_j$ and $f_i$.
      Compute Fundamental matrix using 8-points algorithm.
      Compute Homography matrix using RANSAC.
      Transform points from one image plane to another.
      Compute residual error for both H and F.
      Compute $GRIC_F$ and $GRIC_H$.
      if $GRIC_F$ <$GRIC_H$ then
          $k_j \leftarrow f_i$
      end if
      $i \leftarrow i + 1$
  end while
```

Experiments

- How to check validity of the technique ?
- From 2-4 seconds, video is stationary.
- Algorithm must be able to detect and remove these redundant frames.

Future Directions

- Implementing a few other key-frame selection algorithms.
- Generate labels (Key-frame indexes) for all training videos.
- Neural network training and knowledge transfer between algorithms.

References

[1] Hartley, Richard and Zisserman, Andrew “Multiple View Geometry in Computer Vision”, 2nd Edition, 2004, Cambridge University Press.

[2] Torr P. H. S, 1998, Geometric motion segmentation and model selection, *Philosophical Transactions of Royal Society A,* **356:** 1321–1340.

[3] Ahmed, Mirza Tahir & Dailey, Matthew & Landabaso, José & Herrero, Nicolas. (2010). Robust Key Frame Extraction for 3D Reconstruction from Video Streams.. 1. 231-236.

[4] J. Repko and M. Pollefeys, "3D models from extended uncalibrated video sequences: addressing key-frame selection and projective drift," Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05), 2005, pp. 150-157, doi: 10.1109/3DIM.2005.4.

Thank You!

Any questions?