Building Rome in a Day

Dr. Elli Angelopoulou

Pattern Recognition Lab (Computer Science 5)

University of Erlangen-Nuremberg

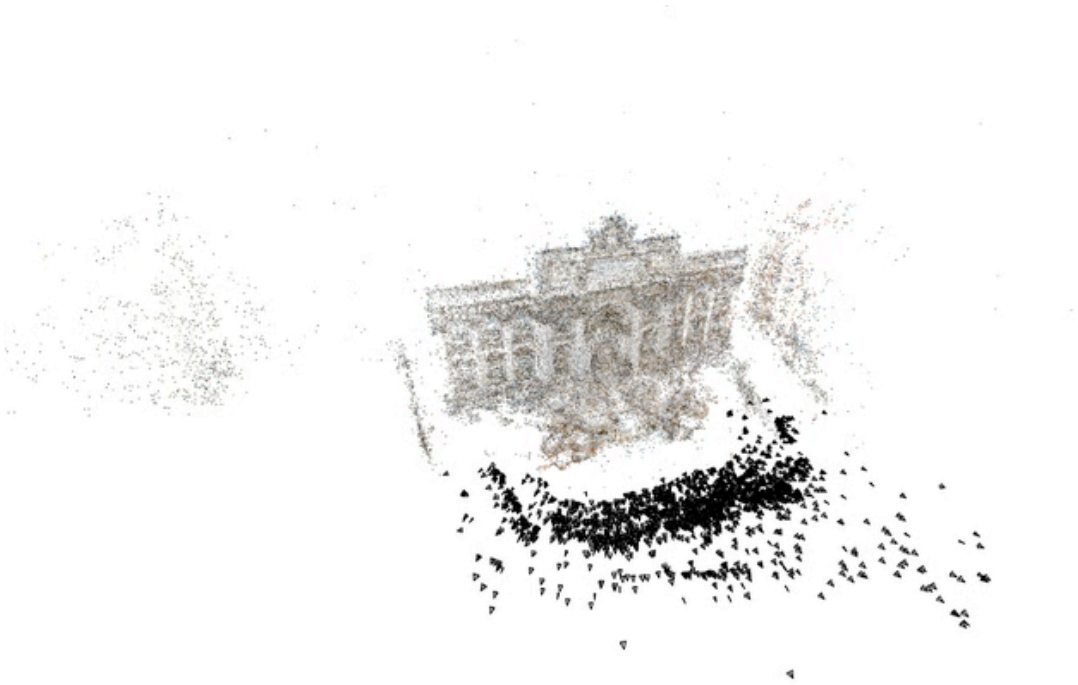
Topic

- A summary of the work of S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. Seitz, R. Szeliski
- Based on the article "Reconstructing Rome", *Computer*, vol. 43, no. 6, June 2010.
- Based also on the article "Building Rome in a Day", *International Conference on Computer Vision* 2010.
- Most of the images and videos in this presentation can be found at http://grail.cs.washington.edu

Idea

- When it comes to famous monuments like the Colosseum in Rome, one can find thousands of images on the web that capture lots of details and many different viewpoints.
- The goal of the project is to use these readily available data to automatically construct a 3D model of the scene.
- Ultimate goal: Build entire city reconstructions within the processing time of a day.
- Advantage: Lots of data readily available. A search for Colosseum in Flickr will return 147,302 results (as of July 19, 2010).
- Challenge 1: Most of the times there is no information about the picture taking process.
  - Unstructured data: pictures are taken in no particular order
  - Uncalibrated data: little is known about the camera type and settings.
- Challenge 2: Too much data to be efficiently processed.

Example of Trevi Fountain



Example of St. Peter's Basilica

Closely Related Work

- This project is still work-in-progress.
- It is part of the Community Photo Collections project at the University of Washington GRAIL Lab.
- Related prior work includes:
  - Photo Tourism (http://phototour.cs.washington.edu/)
  - Skeletal Sets (http://www.cs.washington.edu/homes/snavely/projects/skeletalset/)
  - Photosynth (http://photosynth.net/)

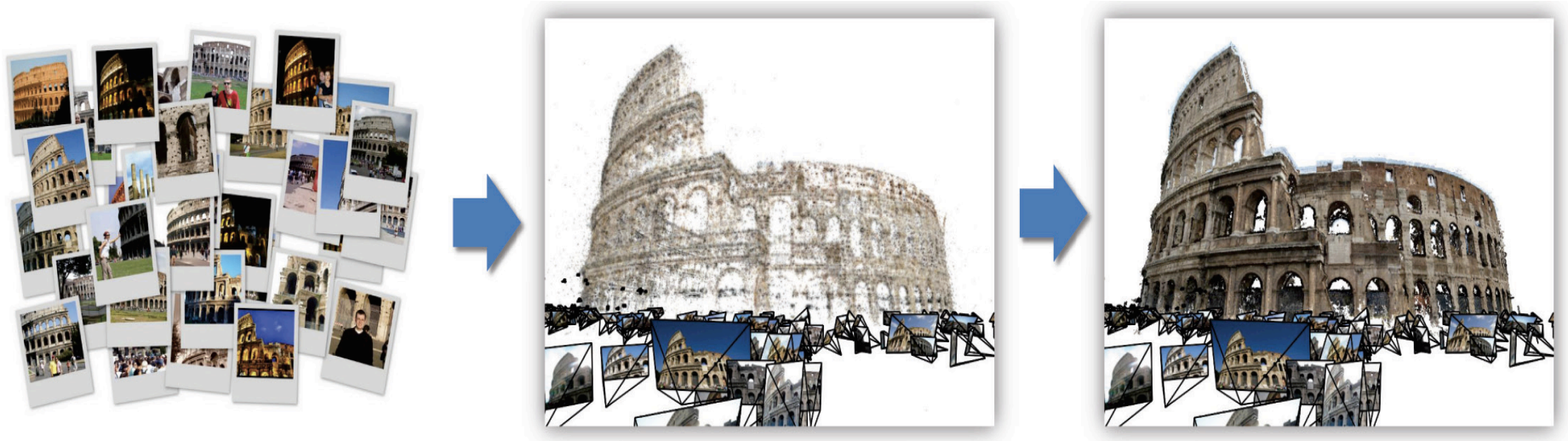
Computer Vision Challenges

- Find common points across images. The descriptors for these characteristic points should be as invariant as possible to changes in viewpoint (rotation, translation, scaling), camera characteristics and illumination.
- Infer camera position and orientation.
- Obtain sparse 3D structure from a set of 2D photos (Structure from Motion, SfM)
- Produce dense 3D geometry from multiple calibrated photos (Multiview Stereo, MVS).

Computer Science Backbone

- Due to its heavy processing demand and very optimistic run-time goals, the "Building Rome in a Day" work relies heavily on:
  - Algorithm Design
  - Distributed Systems
  - Information Retrieval
  - Scientific Computing
  - Computer Graphics

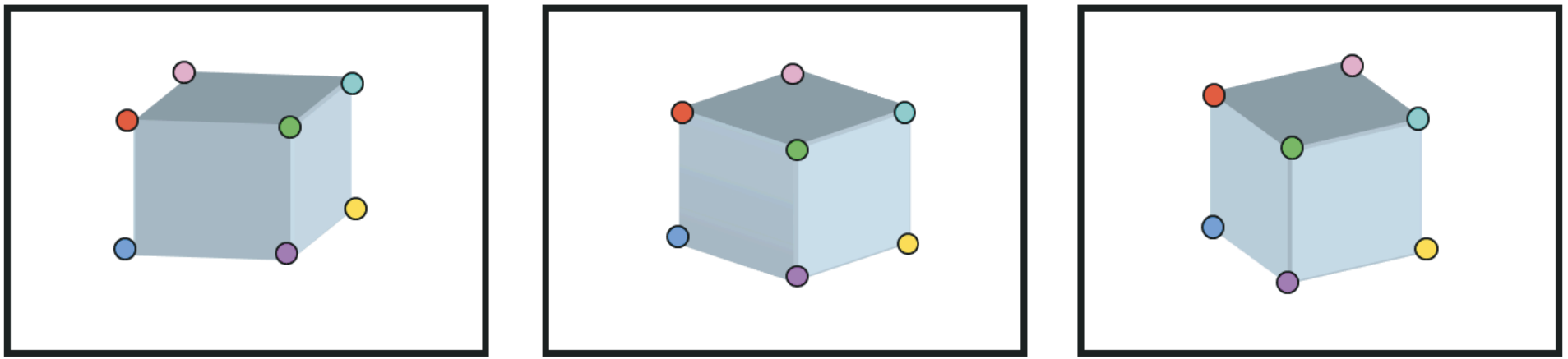
Reconstruction Overview



- Given a set of unstructured, uncalibrated input images.
- First perform SfM to obtain a sparse 3D reconstruction
- Then use MVS to create a dense 3D reconstruction.
- Output: 3D point clouds (working on meshes)
- Running time in 2010: A city (Rome, Dubrovnik, etc.) can be reconstructed from 150K images in less than a day on a cluster with 500 computer cores.

Structure from Motion

- Both the camera viewpoints and the 3D positions of points in the scene are unknown, and must both be solved for simultaneously, using only image data. (Typical SfM setup)
- Start with establishing correspondences.

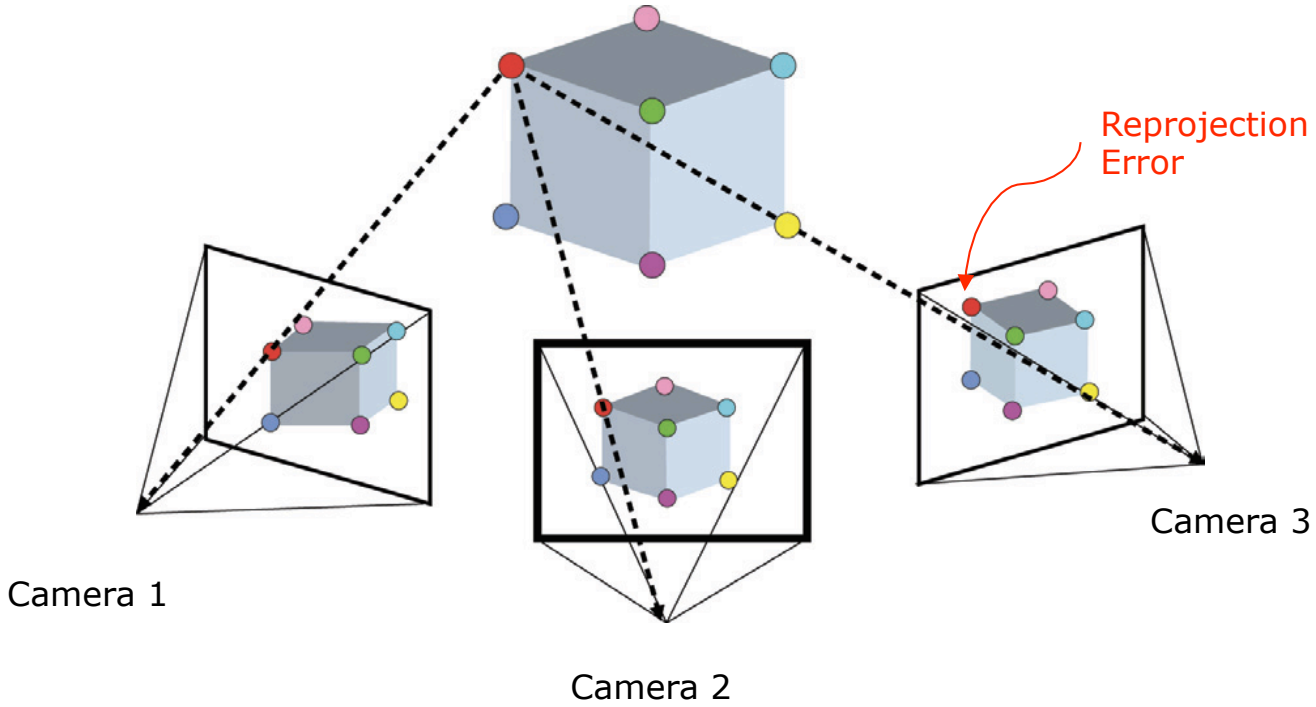


- We don’t know where these images were taken or that they depict a cube. However, let us assume that we know that the corners of the cube as seen in the images are in *correspondence,* that is, each same color triplet depicts the same 3D point.

Structure from Motion – cont.

- The established correspondences provide a powerful constraint on the relative 3D geometry of the cameras and the scene points.
- There is only a limited number of camera-scene configurations that under perspective projection can give rise to these correspondences.
- Goal: Find the camera positions that minimize the sum of the squares of the *reprojection error* (SSRE).

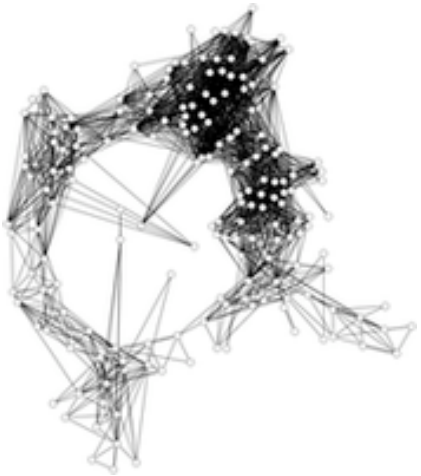
SfM Optimization Problem



- Minimizing the SSRE is a non-linear least squares problem that is difficult to solve:
  - many local minima
  - large scenes have millions of parameters.

SfM Optimization Problem – cont.

- In order to solve this difficult non-linear least squares problem, an incremental approach is used.
- Solve initially for a small number of cameras and points.
- Grow the scene a few cameras at a time. After each round shift around the cameras and points so as to re-optimize the objective function. This process is known as *bundle adjustment* (http://phototour.cs.washington.edu/bundler/)



Recovered Camera Positions

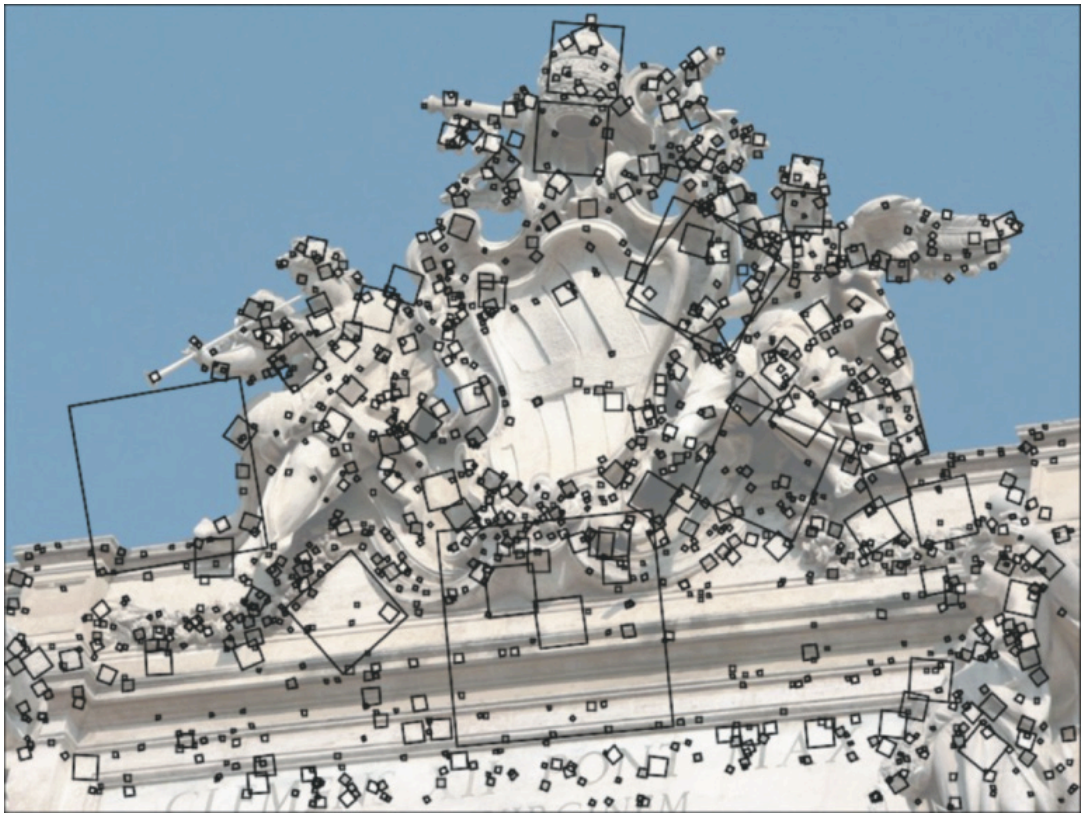
Computed Skeletal Graph

Completed Reconstruction of Stonehenge

Correspondence

- In order to successfully establish correspondences, one needs to identify the most distinctive , repeatable features in an image.

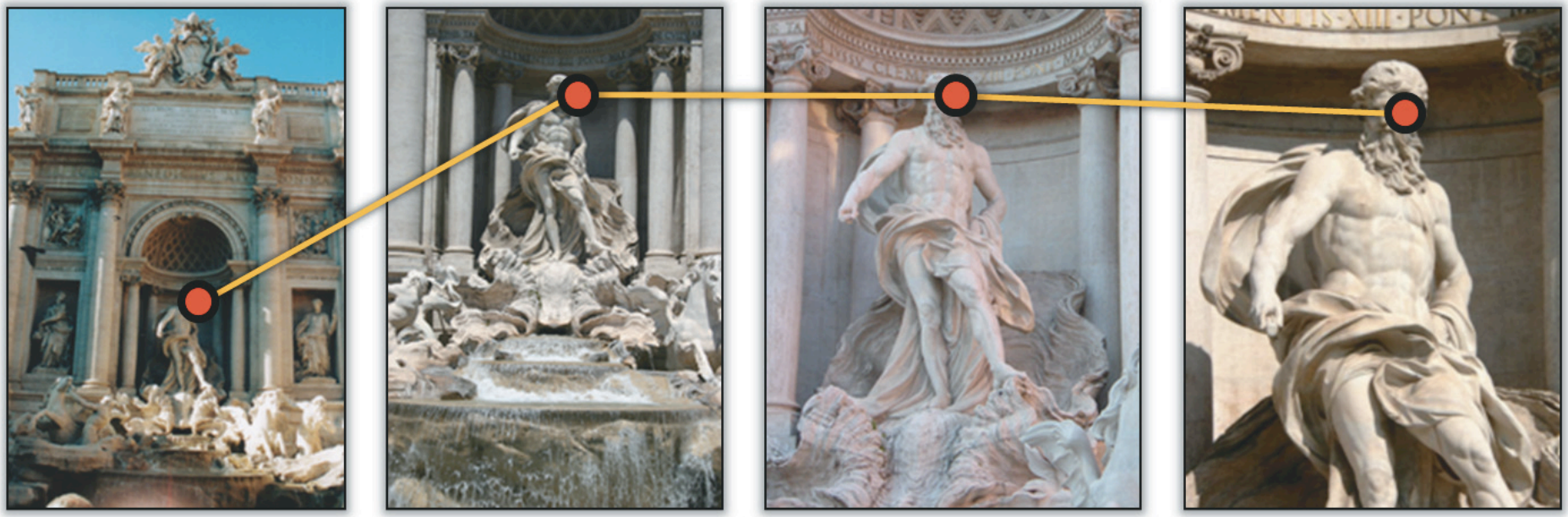
- SIFT features



SIFT features computed on an image of the Trevi Fountain.

Matching Across Image Pairs

- Once features are detected in an image, they can be matched across image pairs by finding similar-looking features (using a variant of the SIFT framework).
- Pairs of matching feature points are linked together to form *tracks* corresponding to the same 3D point in the scene.
- SfM is applied on these tracks of corresponding points.



A track for the face of the central statue of Oceanus at the Trevi Fountain

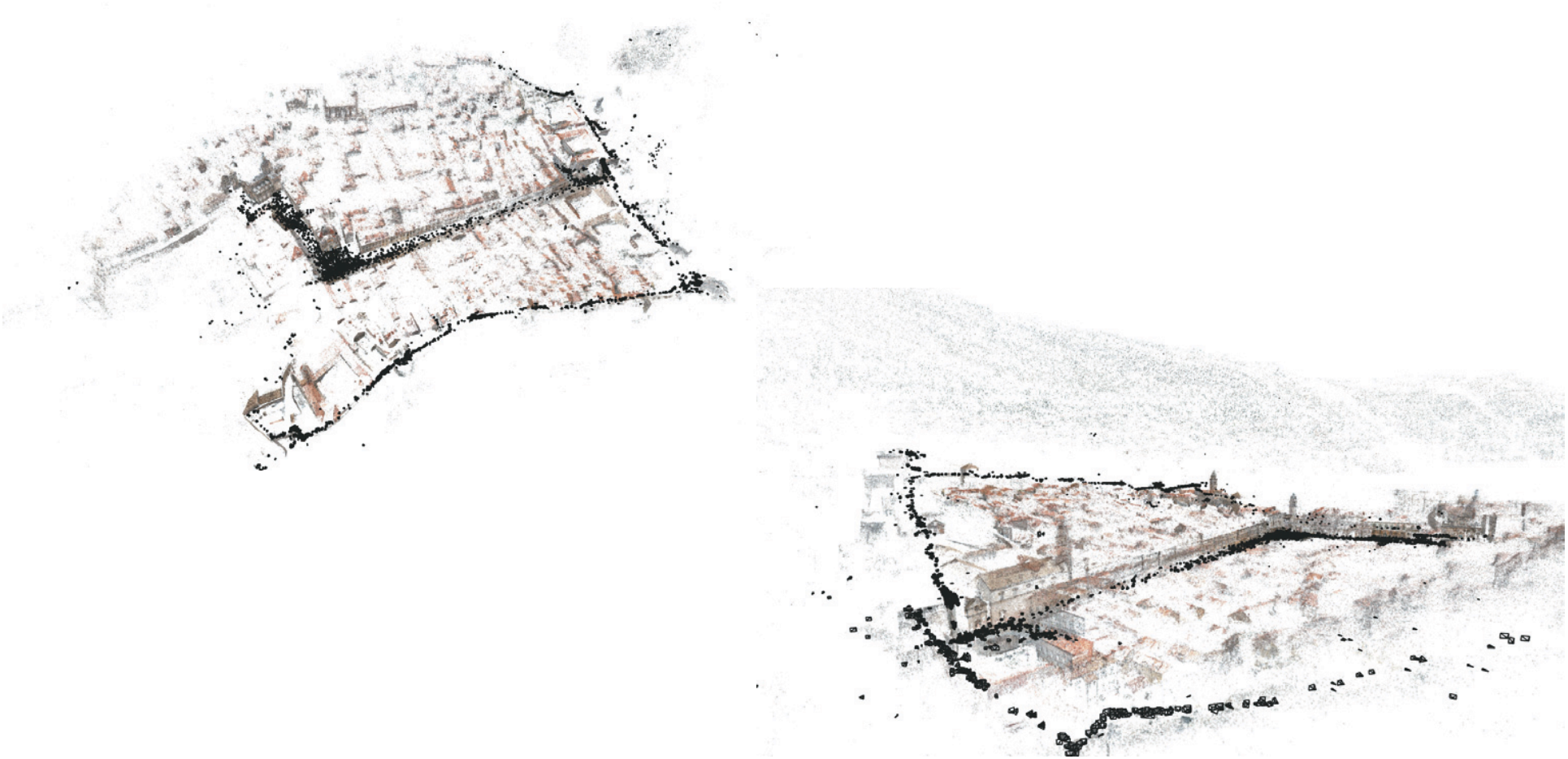
Large Scale Matching

- Comparing pairs of images for computing tracks and then performing SfM can be done for up to a few thousands of images.
- The method doesn't directly scale up to unordered millions of images.
- Solution: Quickly determine pairs of images that look "similar" and then find corresponding points in these image pairs.
- Inspired by work in document analysis, cluster the SIFT features in a photo collection into *visual words*.
- Treat images as documents composed of these visual words, and apply document-retrieval technology (term-frequency analysis and query expansion) to efficiently match large data sets of photos. This is known as the **bag-of-words** technique.

Large-Scale SfM Reconstruction

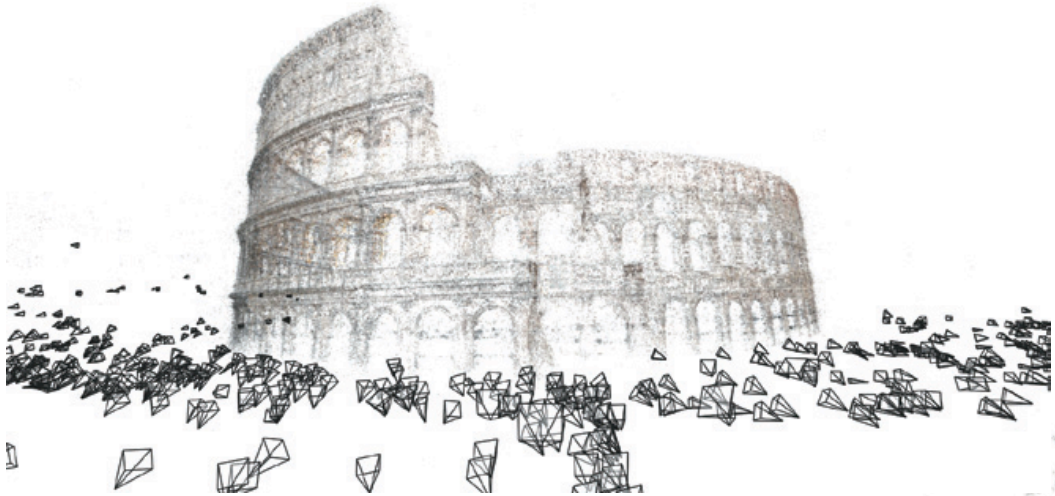
- Many photographs are taken from nearby viewpoints.
- Processing all of them doesn't necessarily add to the reconstruction.
- Find and reconstruct a minimal subset of photos that capture the scene's essential geometry. This process is known as creating the *skeletal graph* of the SfM.
- The remaining images can be considered at a later processing phase, improving performance by an order of magnitude or more.
- By using skeletal graphs reconstructions of:
  - Dubrovnik (58,000 pictures) was performed in 23 hrs (5hrs for matching 18hrs for reconstruction) using 352 processors.
  - Rome (150,000 pictures) was performed in 26 hrs (18hrs for matching 8hrs for reconstruction) using 496 processors.
  - Venice (250,000 pictures) was performed in 65 hrs (27hrs for matching 38hrs for reconstruction) using 496 processors.

Dubrovnik Reconstruction



- Reconstructing Dubrovnik (58,000 pictures) took 23 hrs (5hrs for matching 18hrs for reconstruction) using 352 processors.

Rome Reconstruction

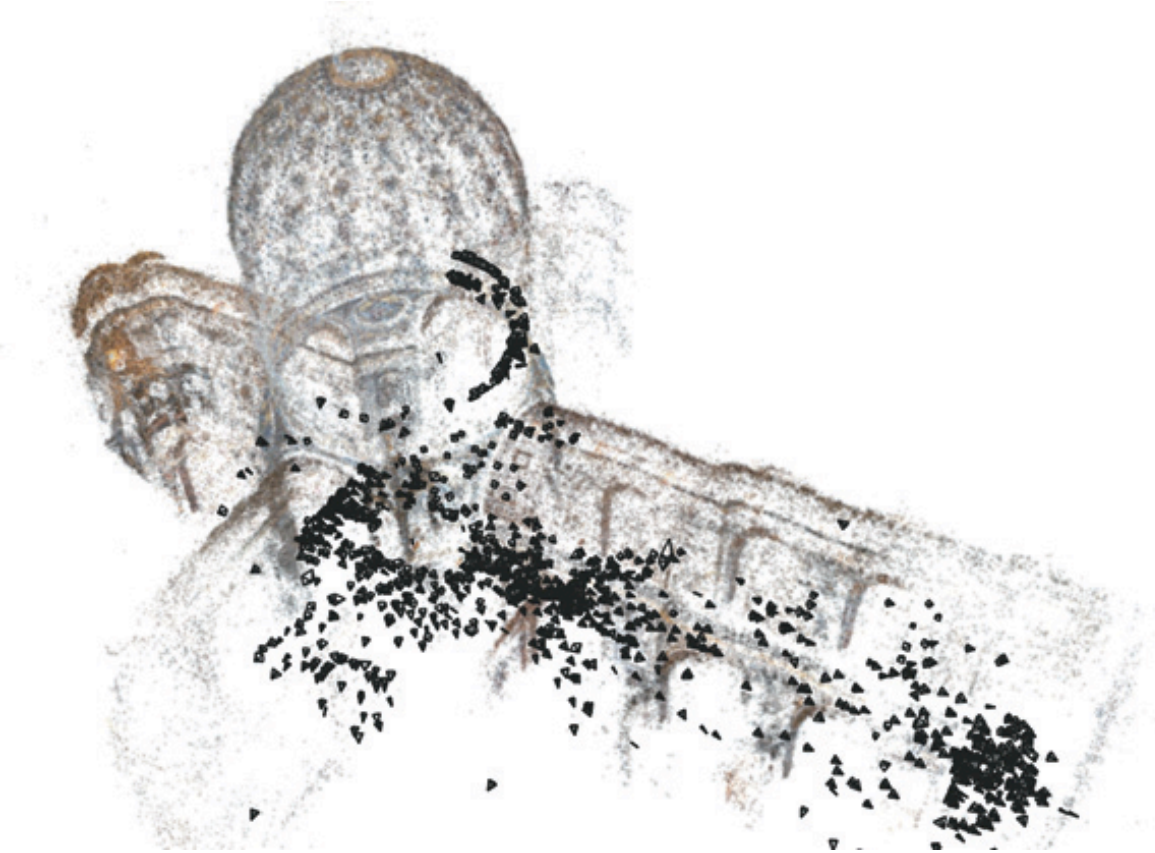


Colosseum – 2,106 photos

Trevi Fountain – 1,936 photos

- Reconstructing Rome (150,000 pictures) took 26 hrs (18hrs for matching 8hrs for reconstruction) using 496 processors .

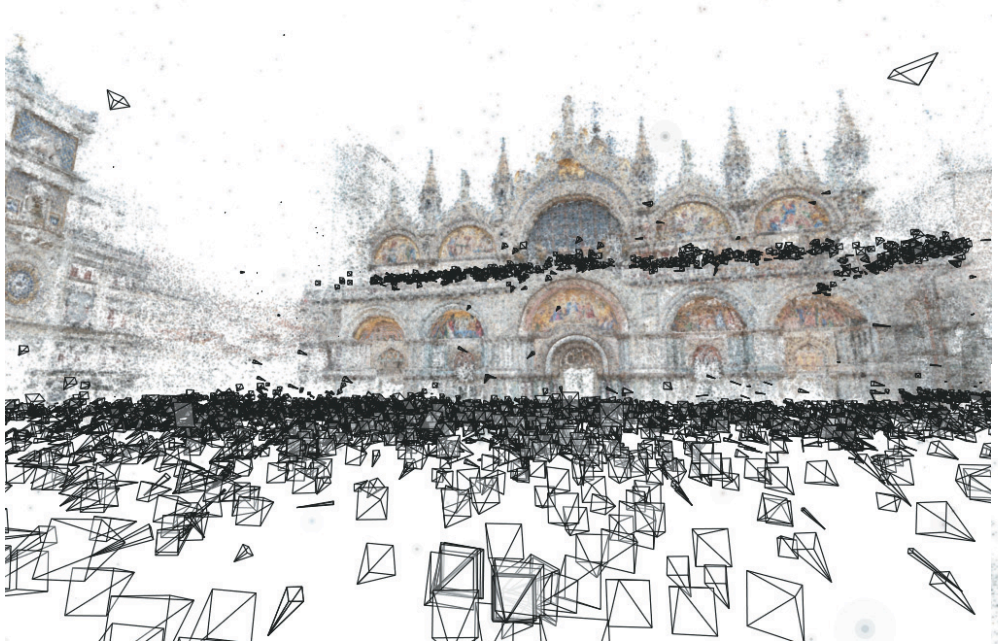
Rome Reconstruction



St. Peter's Basilica – 1,294 photos

- Reconstructing Rome (150,000 pictures) took 26 hrs (18hrs for matching 8hrs for reconstruction) using 496 processors .

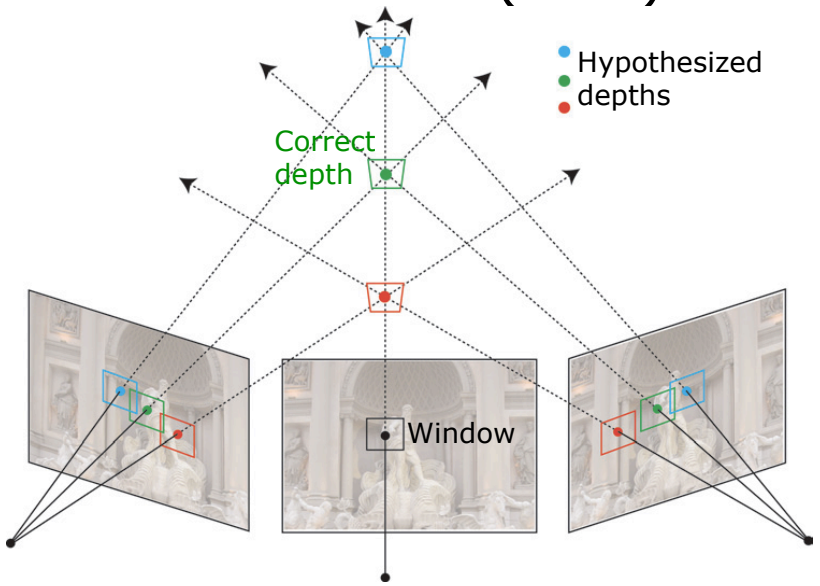
Venice Reconstruction



- The reconstruction of Venice was based on 250,000 pictures. It took 65 hrs (27hrs for matching 38hrs for reconstruction) using 496 processors. Shown here is the reconstruction of the San Marco Plaza using 13,699 images.

Dense 3D Reconstruction

- SIFT features are distinctive and repeatable and allow for large-scale SfM reconstruction.
- However, 3D data can only be recovered at SIFT keypoints, leading to a sparse 3D reconstruction.
- Once images are registered against each other (i.e. we know the relative position - rotation and translation – of the cameras that took the photos) one can use multiview stereo (MVS).
- The proposed system uses a patch based consistency-check MVS algorithm.
- A depth estimate is computed for every pixel.



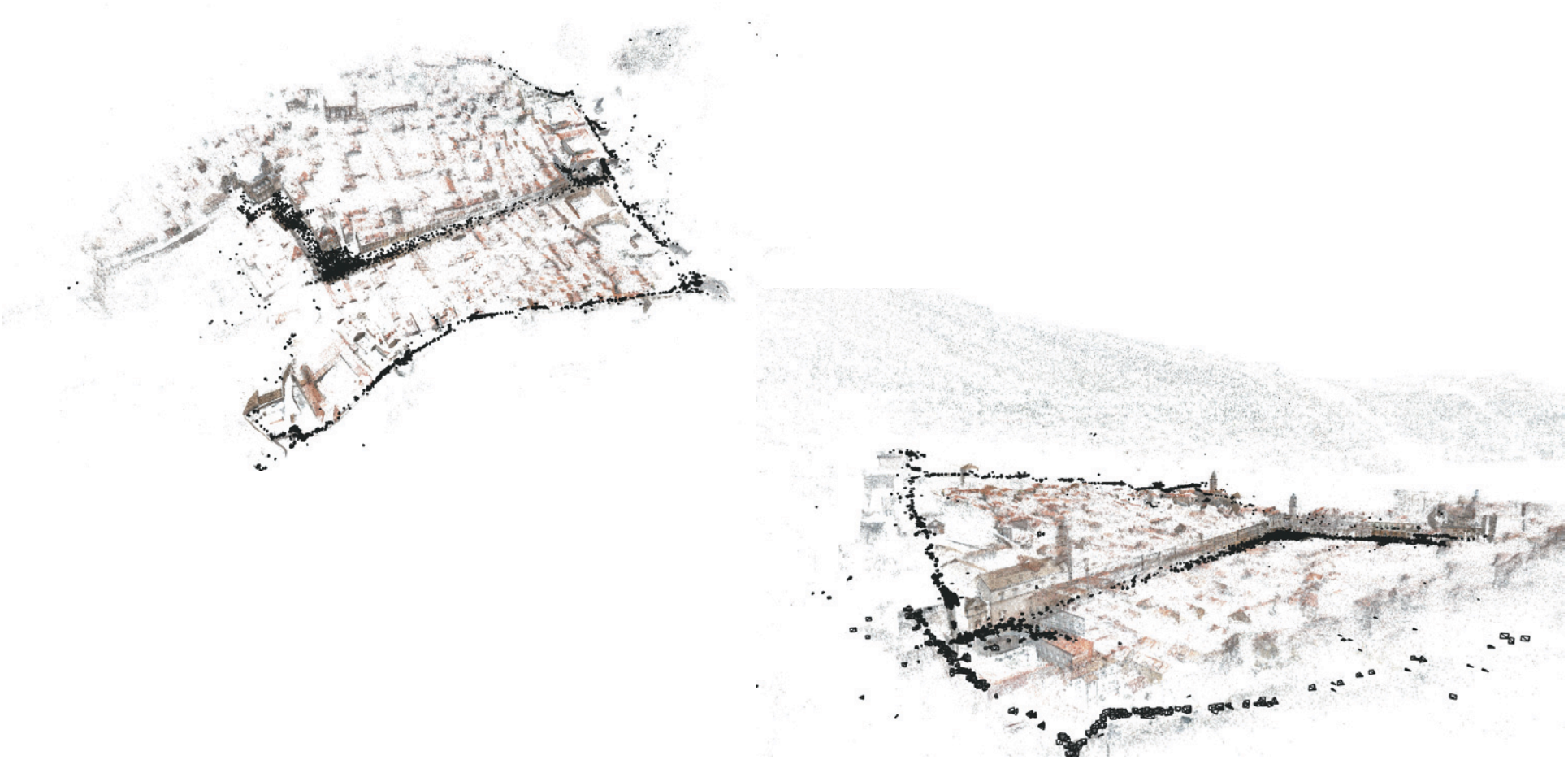
Large-Scale MVS

- Standard stereo algorithms can not process these thousands of images all at once (memory limitations).
- Group photos into manageable-sized clusters, and calculate scene geometry using MVS within each cluster independently.
- How to cluster? Select images that are looking at the same part of the scene. (Information available through the SfM analysis)
- Problem: The reconstruction from a cluster consisting of very similar viewpoints would be poor.
- Solution: Remove redundant images that are very close, or even identical, to other viewpoints.
- The final image clustering for MVS is also expressed as an optimization problem: use a sufficiently small cluster size and a relatively complete scene coverage, while minimizing the total sum of the cluster sizes.

Large-Scale MVS – cont.

- Once the clusters are formed, MVS is run independently on each cluster.
- This can be done in parallel, thus speeding up the dense reconstruction.
- The output of each MVS-cluster is a point cloud.
- The point clouds are merged into a single model, a single large consistent point cloud for the entire scene (city).
- There is ongoing work on efficiently generated meshes out of the point clouds.

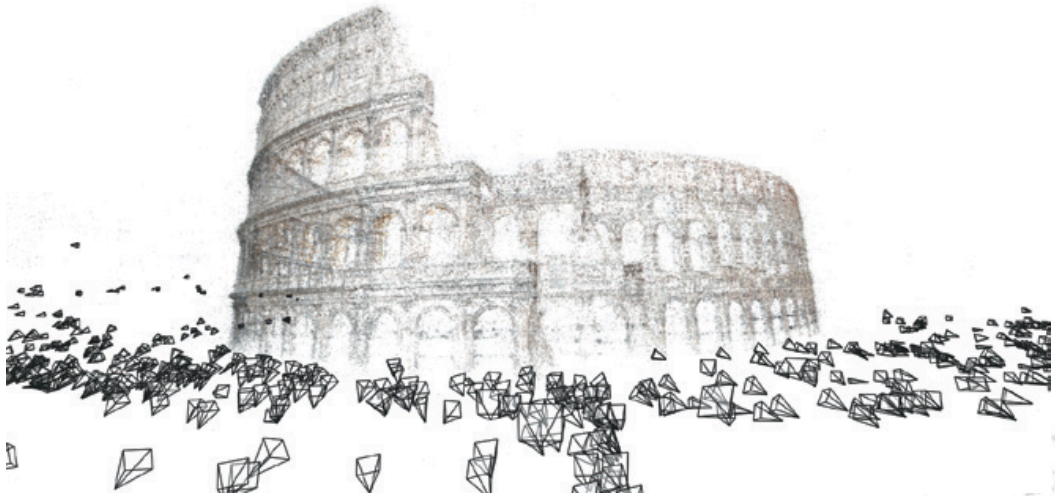
Dubrovnik Reconstruction



- Reconstructing Dubrovnik (58,000 pictures) took 23 hrs (5hrs for matching 18hrs for reconstruction) using 352 processors.

Dense Dubrovnik Reconstruction

Rome Reconstruction



Colosseum – 2,106 photos

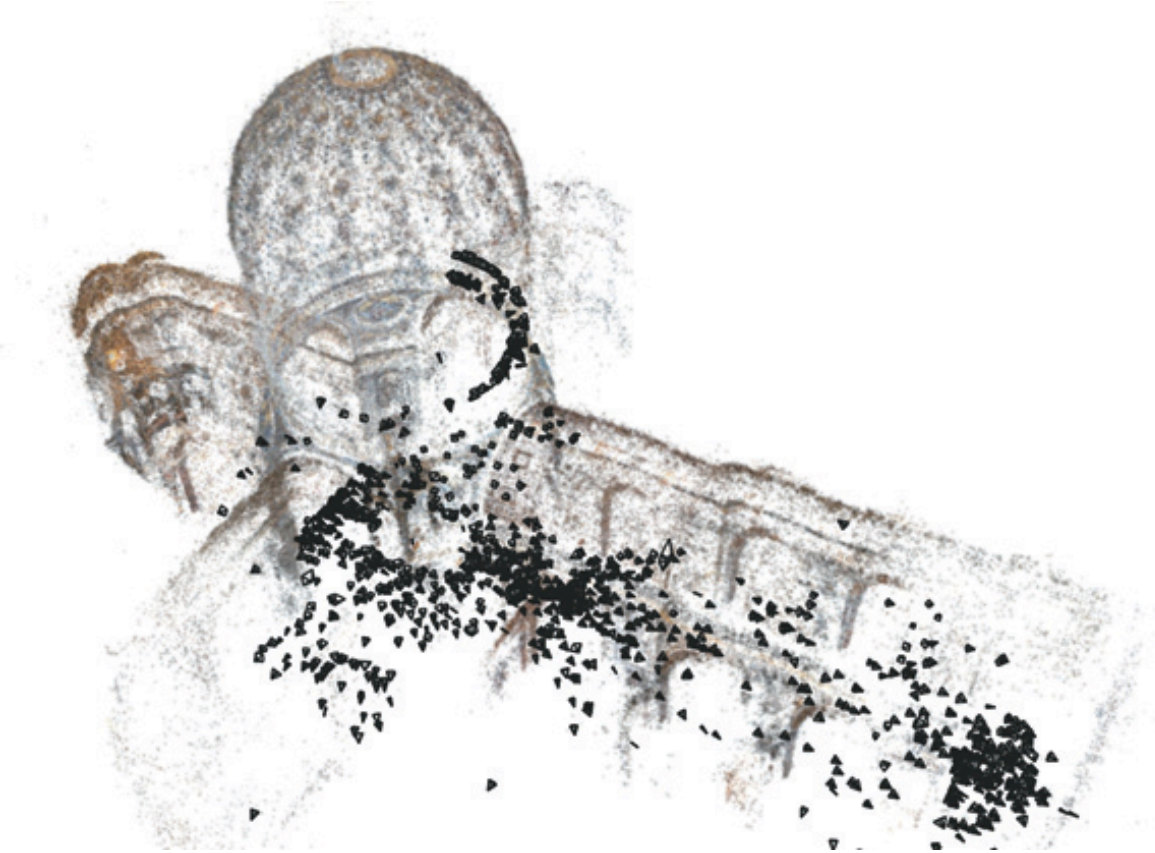
Trevi Fountain – 1,936 photos

- Reconstructing Rome (150,000 pictures) took 26 hrs (18hrs for matching 8hrs for reconstruction) using 496 processors .

Dense Colosseum Reconstruction

Dense Colosseum Reconstruction

Rome Reconstruction



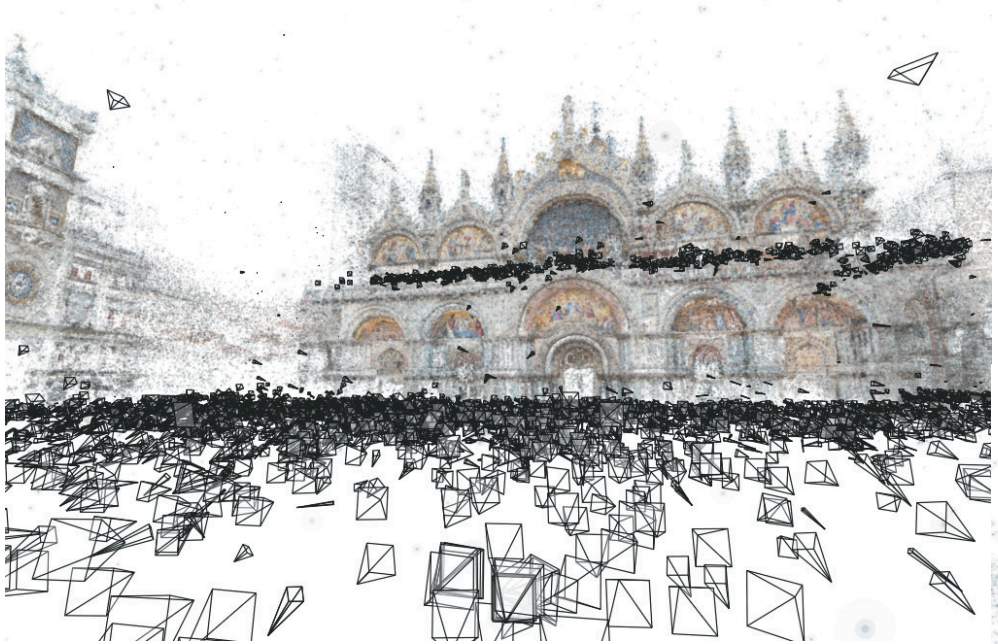
St. Peter's Basilica – 1,294 photos

- Reconstructing Rome (150,000 pictures) took 26 hrs (18hrs for matching 8hrs for reconstruction) using 496 processors .

Dense St. Peter's Basilica Reconstruction

Dense St. Peter's Basilica Reconstruction

Venice Reconstruction



- The reconstruction of Venice was based on 250,000 pictures. Shown here is the reconstruction of the San Marco Plaza using 13,699 images.

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza

Dense Reconstruction of San Marco Plaza