Large-scale 3D Modeling from Crowdsourced Data

Johannes Schönberger, ETHZ
Jared Heinly, URCV
Enrique Dunn, SIT
Jan-Michael Frahm, UNC
Marc Pollefeys, Microsoft, ETHZ

URCV ETH zürich Microsoft

People

- Jan-Michael Frahm, UNC Chapel Hill
- Marc Pollefeys, Microsoft, ETH-Zürich

- Johannes Schönberger, ETH-Zürich

- Jared Heinly, URCV

- Enrique Dunn, Stevens Institute of Technology

A World of Cameras

- Close to a ***quadrillion*** photos taken last year
- ***Trillions*** uploaded every year

Super Sensor

Diverse

Uncontrolled

Asynchronous

100 Million Images Yahoo

Camera


100 million images
Frankfurtdoku
Dakil
Lot # 223-224
FINISH

Visual Index of the World

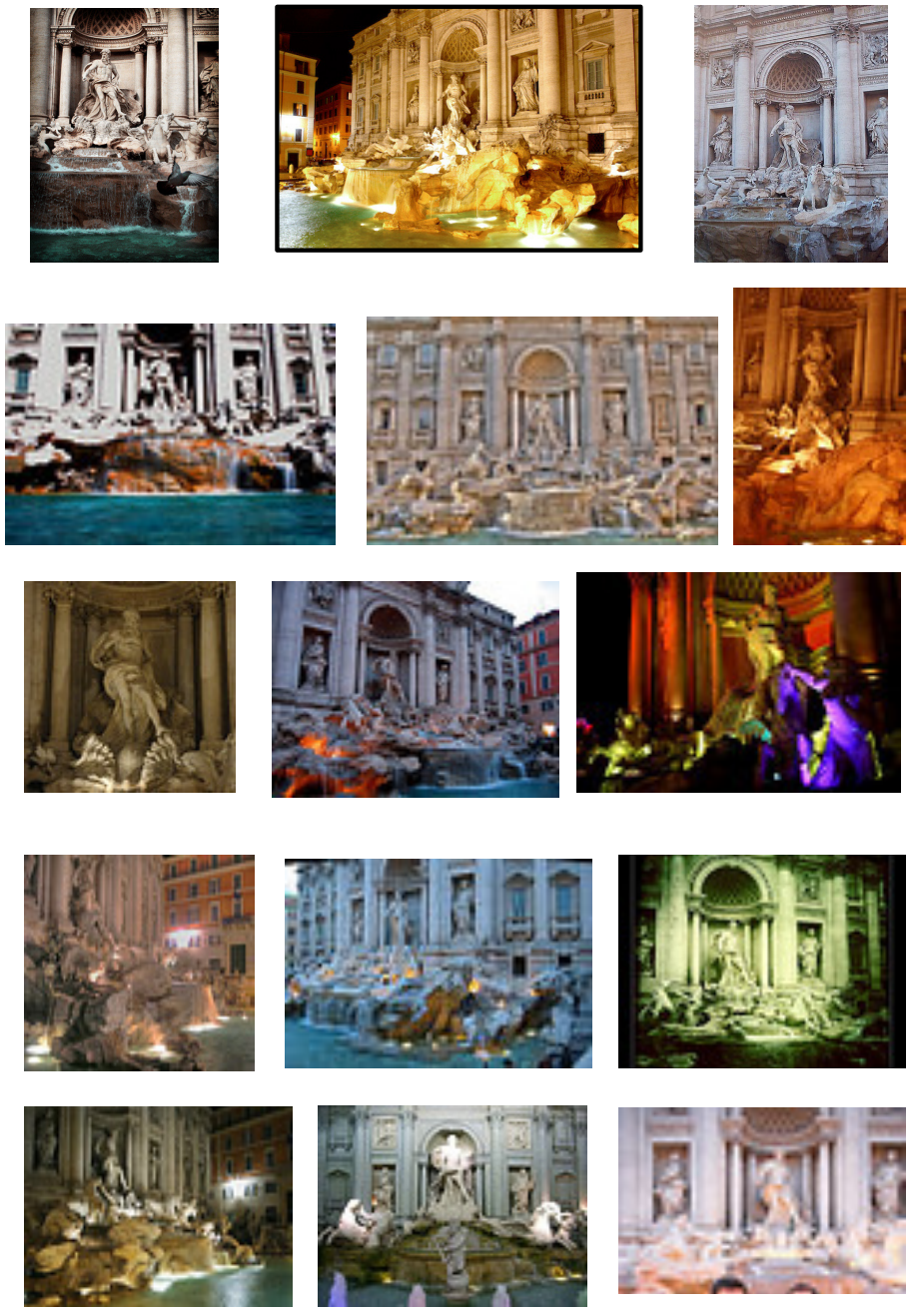
Virtual World Model

Challenges of Large-scale Reconstruction

- Robustness
- Scalability
- Completeness

Challenges for Reconstruction

- Robustness
- Scalability
- Completeness
- Appearance Variation



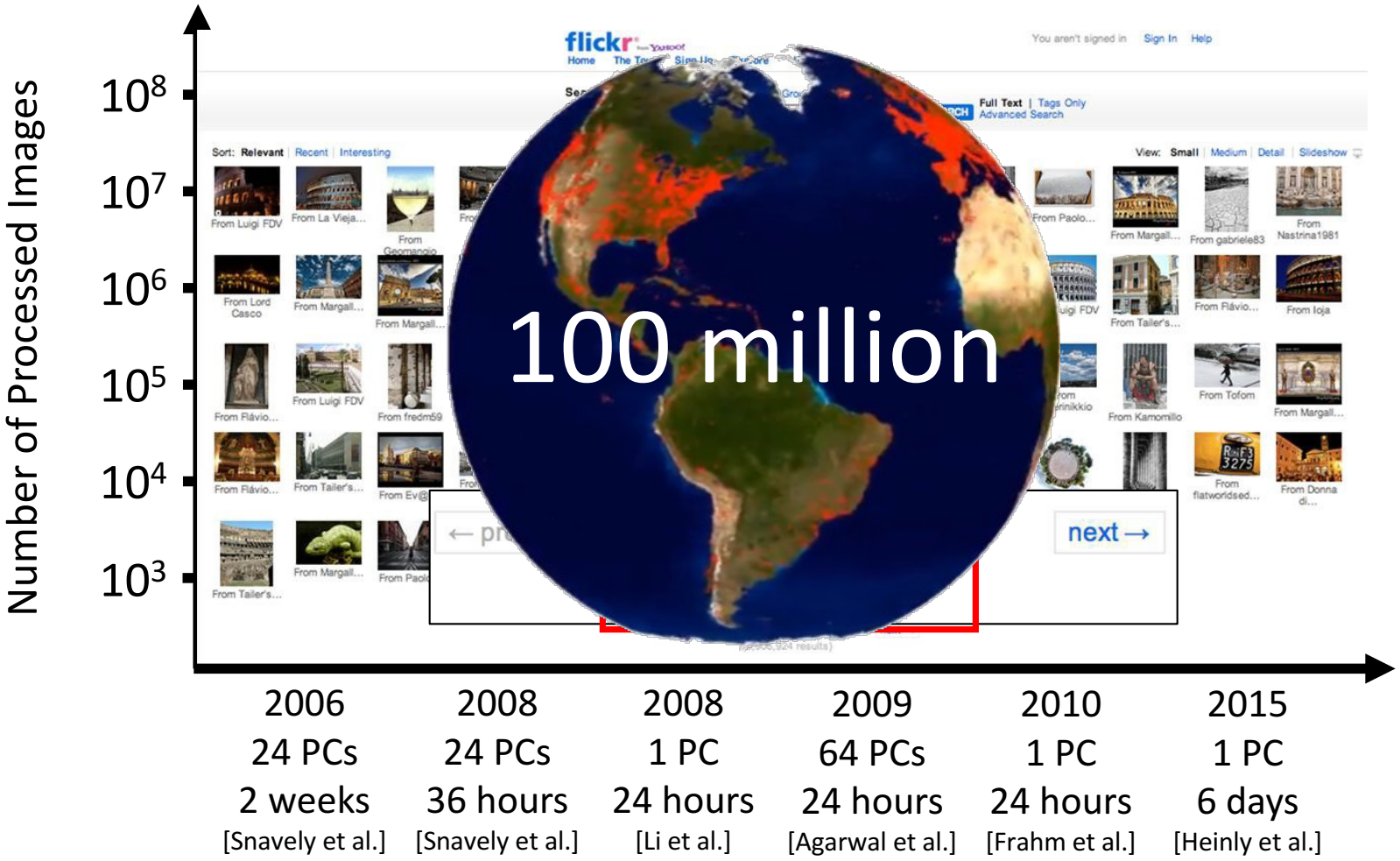
URCV ETH zürich Microsoft

Challenges for Reconstruction

- Robustness
- Scalability
- Completeness
- Appearance Variation
- Surface Resolution

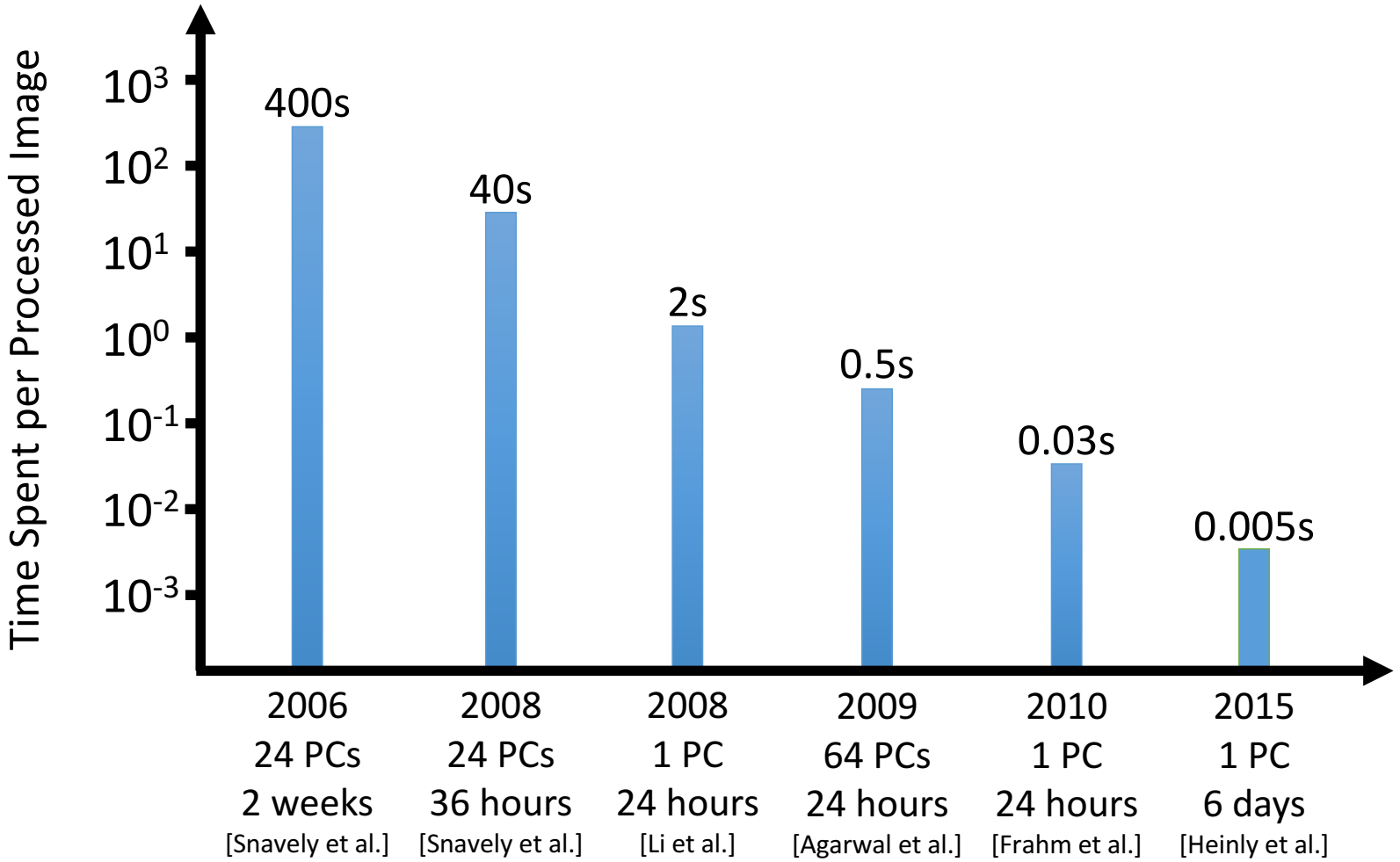
URCV ETH zürich Microsoft

Large-Scale Crowd-Sourced 3D Modeling



URCV ETH zürich Microsoft

Large-Scale Crowd-Sourced 3D Modeling



Problem Statement

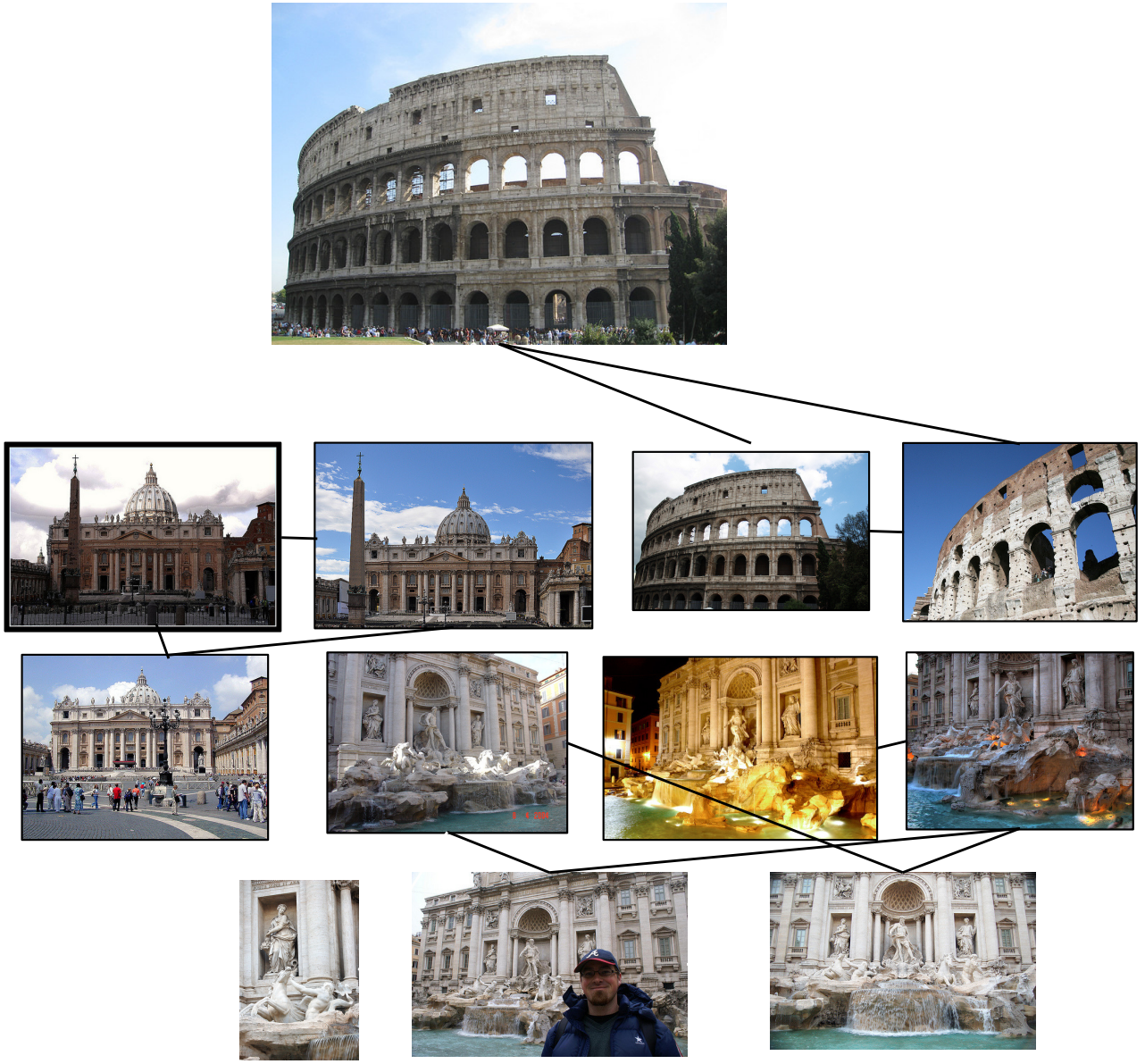
Recover the 3D geometry and appearance of the 3D scene from 2D photographs/videos captured from multiple viewpoints.

Data Association

Data Association

images, 2D features

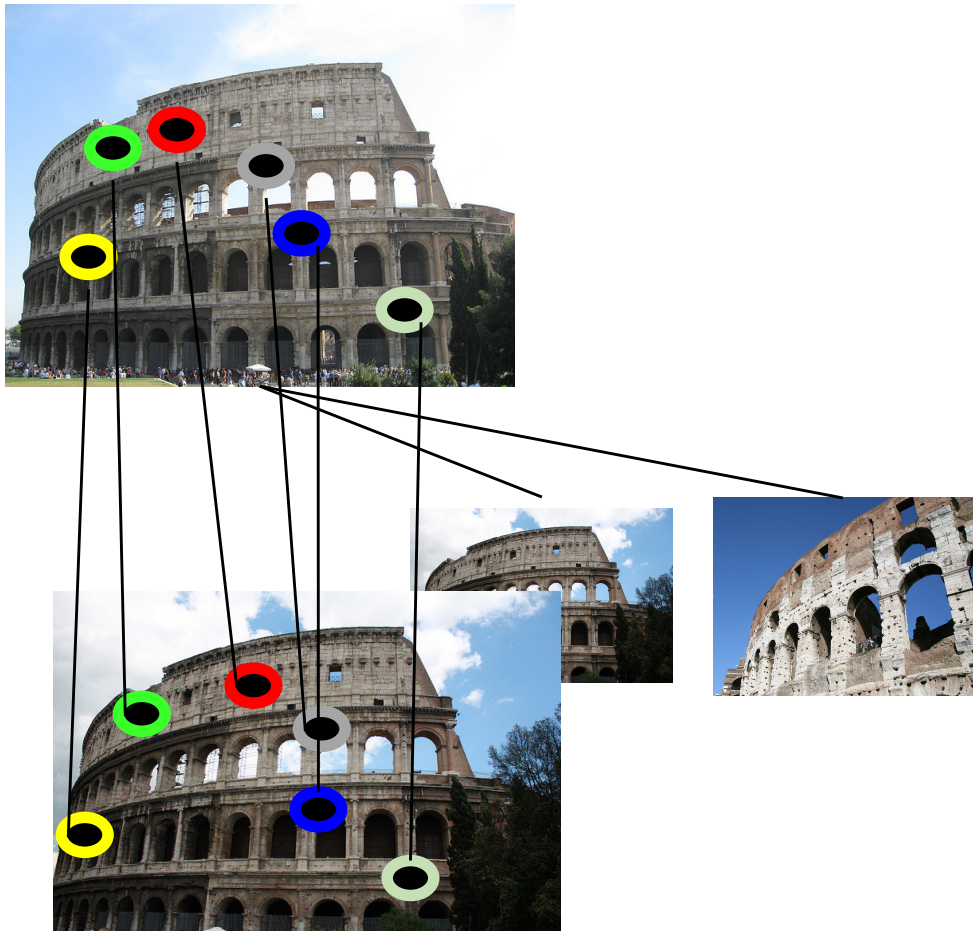
Jared Heinly



Spatial Correlation

Data Association

images, 2D features
correspondences



3D from images

Data Association

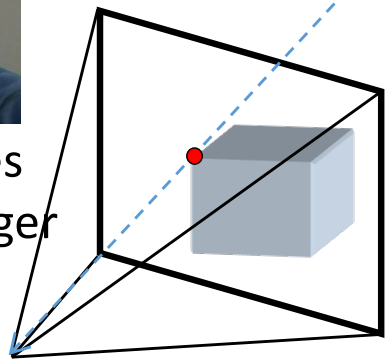
images, 2D feature correspondences

3D points & camera poses

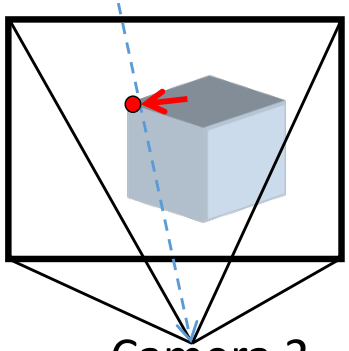
images, camera poses

Marc Pollefeys

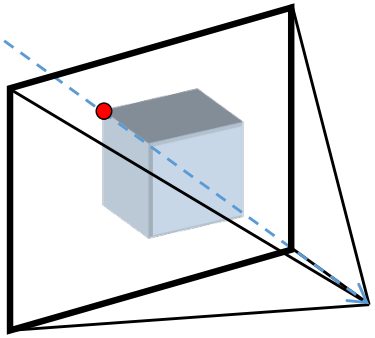
Johannes Schönberger



Camera 1



Camera 2



Camera 3

Dense 3D from images

Data Association

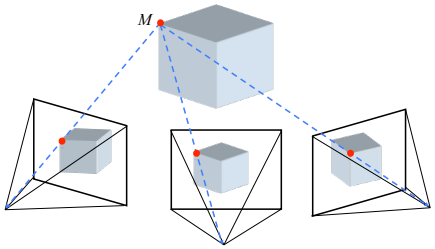
images, 2D feature correspondences

3D points & camera poses

images, camera poses

Dense 3D

images, camera poses, depth maps



Enrique Dunn

Dense 3D from images

Data Association

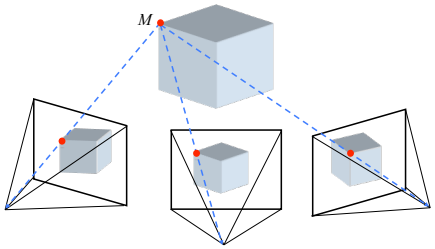
images, 2D feature correspondences

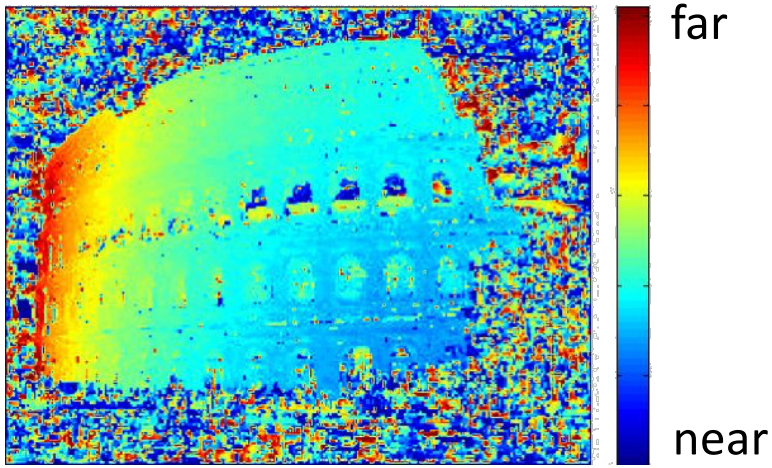
3D points & camera poses

images, camera poses

Dense 3D

images, camera poses, depth maps





Dense 3D from images

Data Association

images, 2D feature correspondences

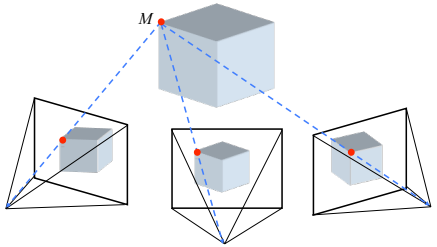
3D points & camera poses

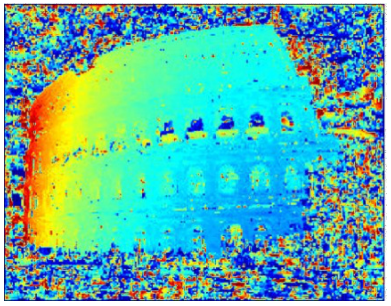
images, camera poses

Dense 3D

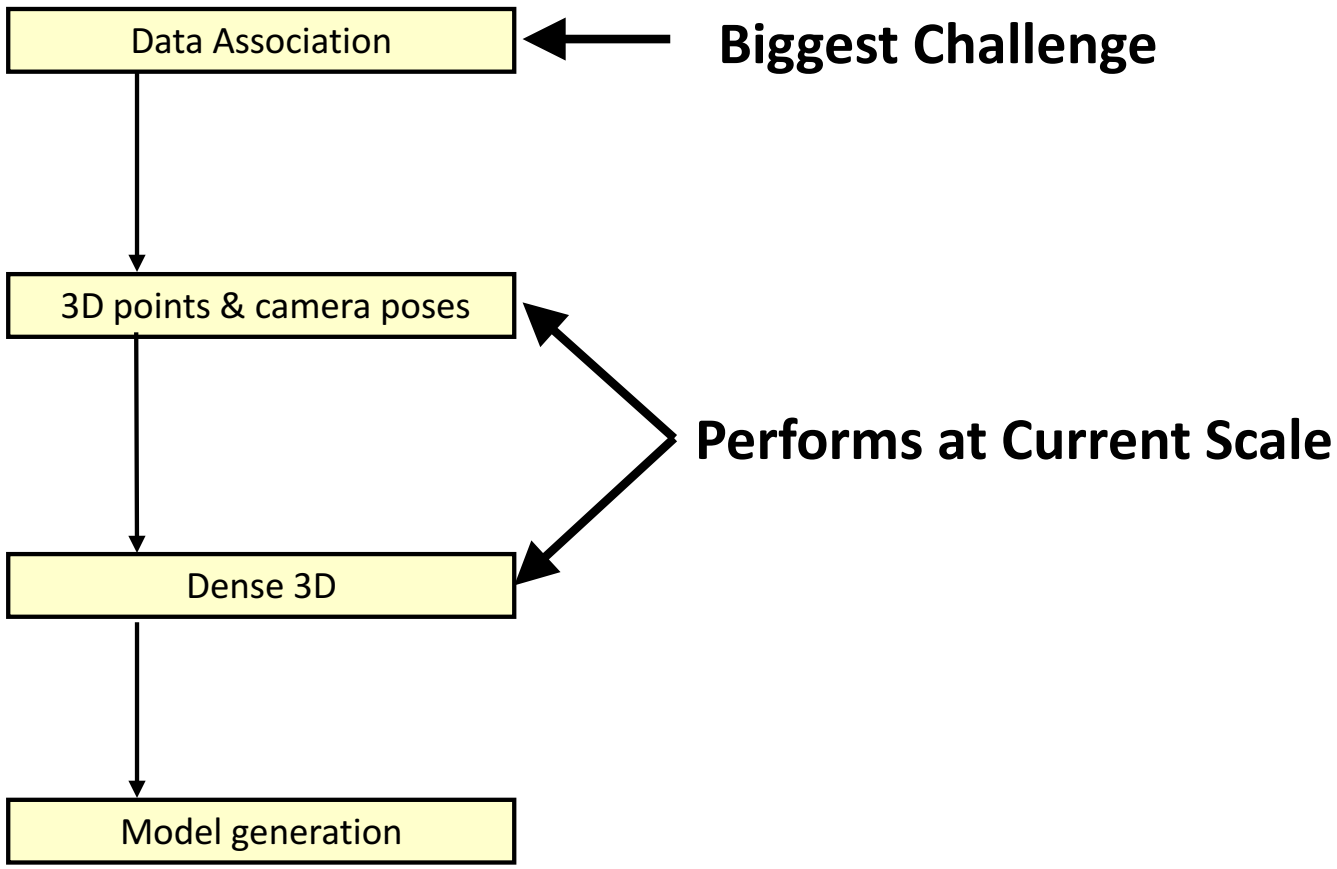
images, camera poses, depth maps

Model generation





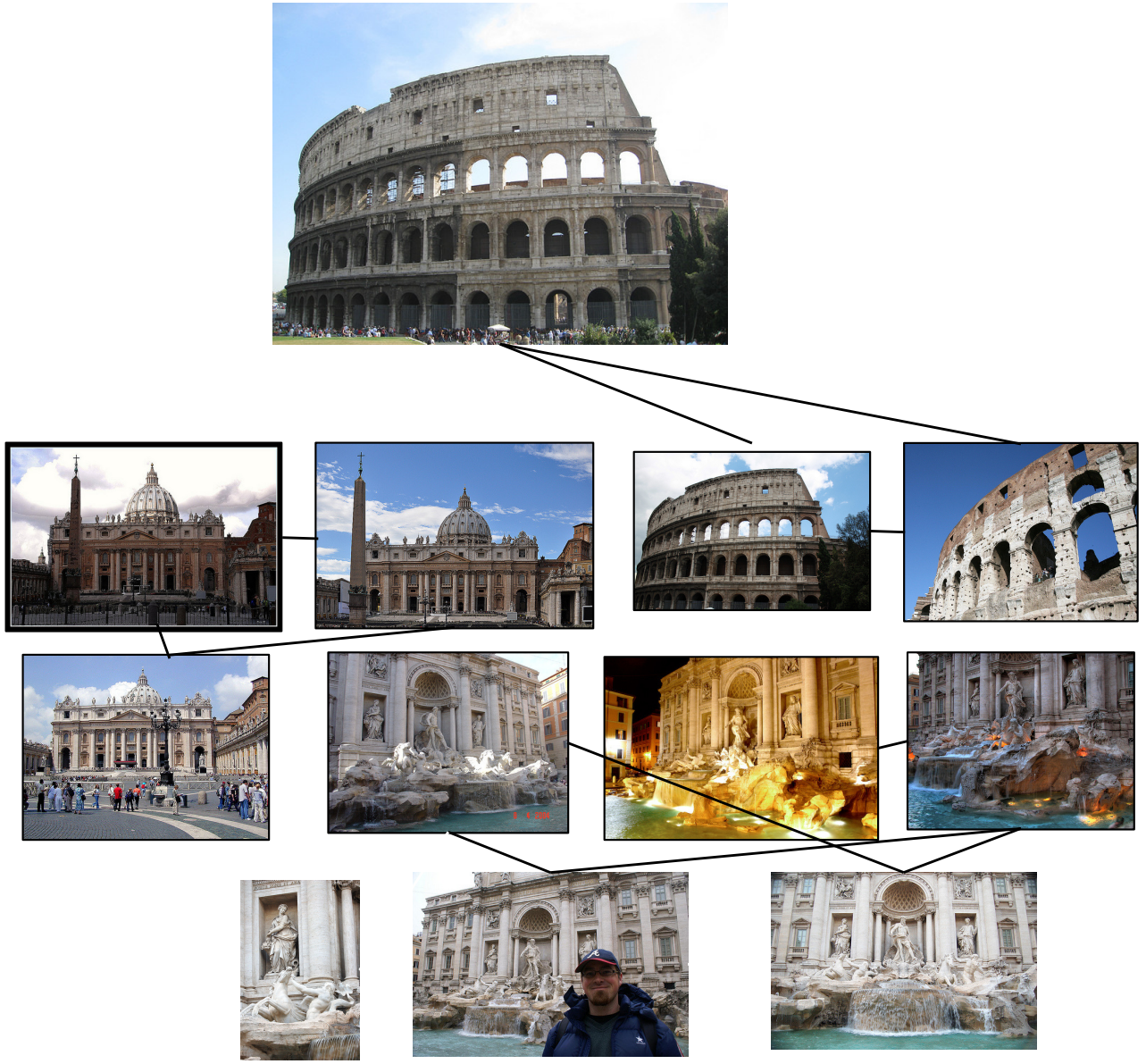
3D Modeling Pipeline



Data Association

Data Association

images, 2D features



Data Association

Data Association

images, 2D features

Image correspondence

3D from images

Data Association

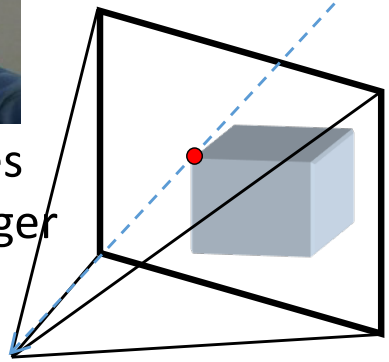
images, 2D feature correspondences

3D points & camera poses

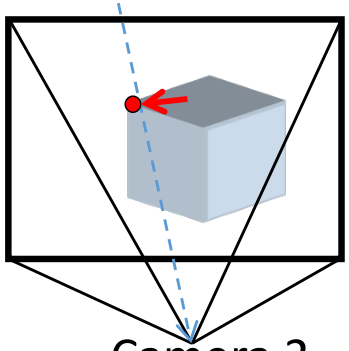
images, camera poses

Marc Pollefeys

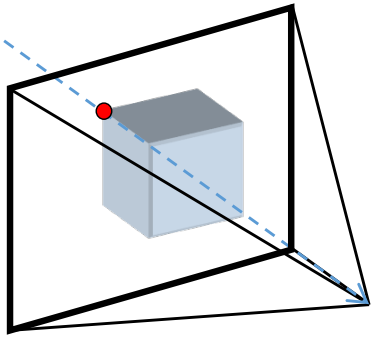
Johannes Schönberger



Camera 1



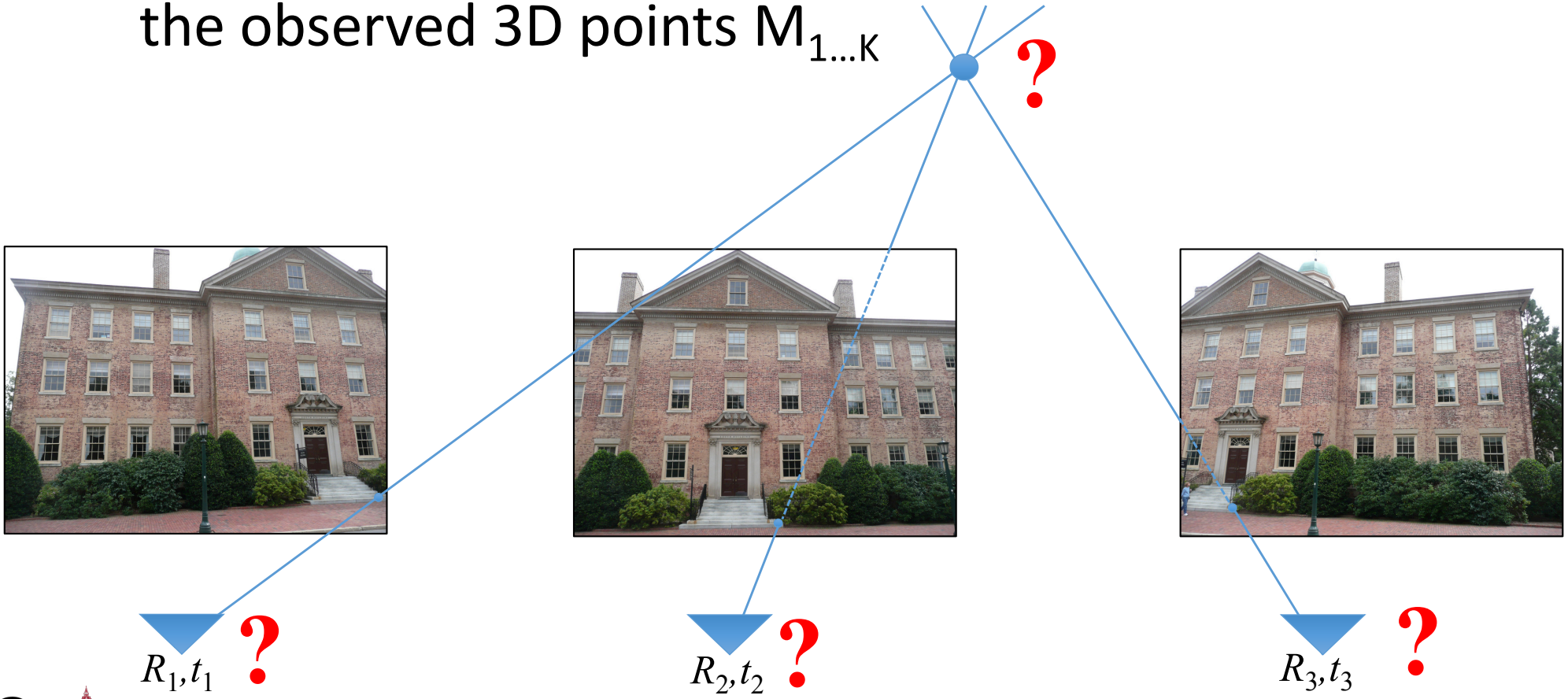
Camera 2



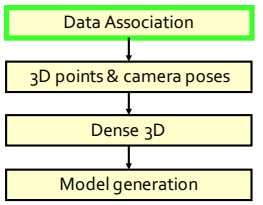
Camera 3

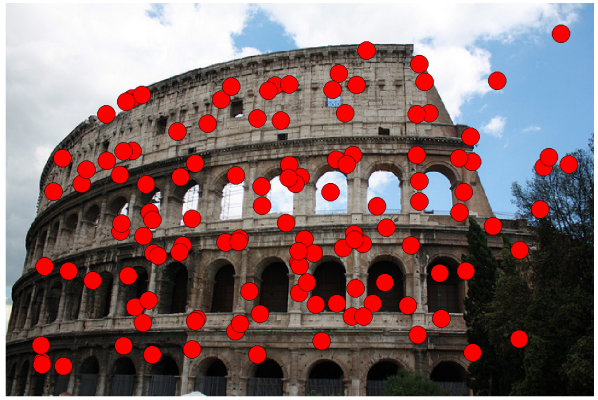
Structure from Motion

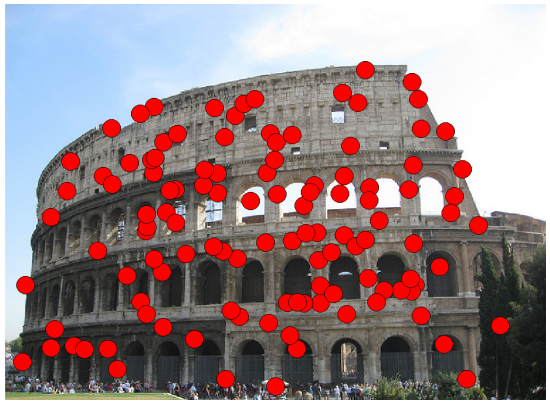
- Given a set of ***N*** images, observing ***K*** static 3D points $m_{i,j} = P_i M_j$ with i=1,…,N, j=1,…,K
- Recover the ***N*** camera projection matrices $P_{1…N}$ and the observed 3D points $M_{1…K}$



Geometric verification







- RANSAC (computational cost is exponential in outlier fraction)
- typically < 20 Hz computation with traditional methods
- Faster computation [“ARRSAC”, ECCV 2008],
- Error propagation for better performance [“Cov-RANSAC”, ICCV 2009]
- Generalized RANSAC [“USAC”, PAMI 2013]

~625 Hz geometric verification at same quality


URCV ETH zürich Microsoft

Dense 3D from images

Data Association

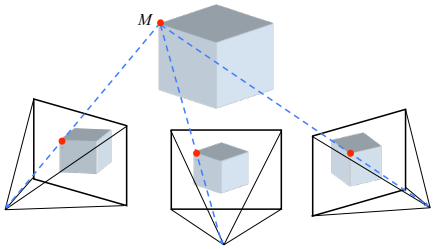
images, 2D feature correspondences

3D points & camera poses

images, camera poses

Dense 3D

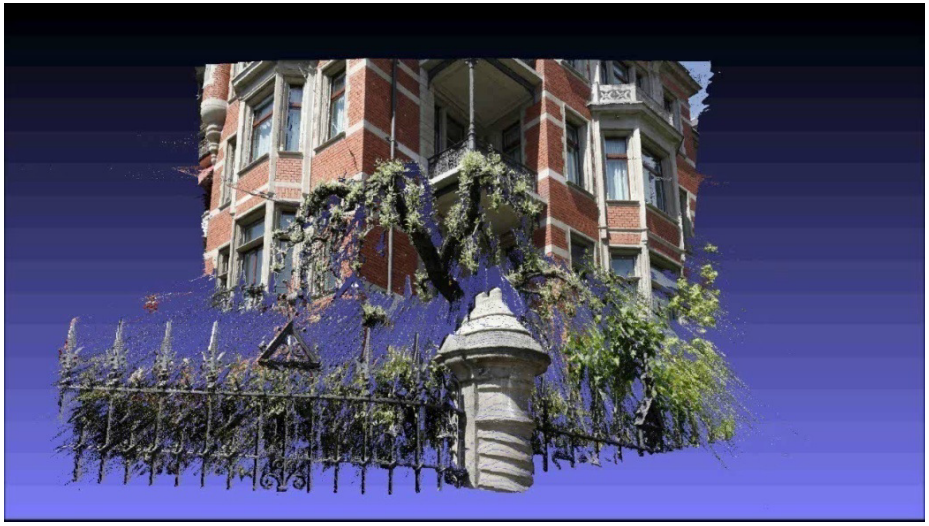
images, camera poses, depth maps



Enrique Dunn

Stereo

- Extraction of 3D information from 2D images



Images → Stereo → 3D Point Cloud

Video credit: A. Hornung

Available Structure from Motion Software

- Bundler https://www.cs.cornell.edu/~snavely/bundler
- VisualSFM http://ccwu.me/vsfm (output to dense)
- COLMAP https://colmap.github.io (includes dense)
- Theia http://www.theia-sfm.org

Data association

- Streaming Association https://github.com/jheinly/streaming_connected_component_discovery

Large-scale SFM Papers (selection)

- S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, “Bundle adjustment in the large.” ECCV 2010
- S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. “Building Rome in a Day.” Comm. ACM, 2011
- J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. “Building Rome on a cloudless day.” ECCV 2010
- J. Heinly, J. L. Schönberger, E. Dunn, J.-M. Frahm, “Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset)“, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
- X. Li, C. Wu, C. Zach, S. Lazebnik, J.-M. Frahm. “Modeling and Recognition of Landmark Image“. ECCV 2008
- R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik. “Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs.” IJCV, 2011
- N. Snavely, S. M. Seitz, and R. Szeliski. “Photo Tourism: Exploring image collections in 3D.” SIGGRAPH, 2006
- [N.Snavely, S. M Seitz, and R. Szeliski. “Modeling the world from internet photo collections.” IJCV, 2008
- N. Snavely, S. M Seitz, and R. Szeliski. “Skeletal graphs for efficient structure from motion.” In CVPR, 2008.

Data Association

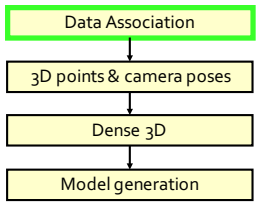
Data Association

images, 2D features

Image correspondence

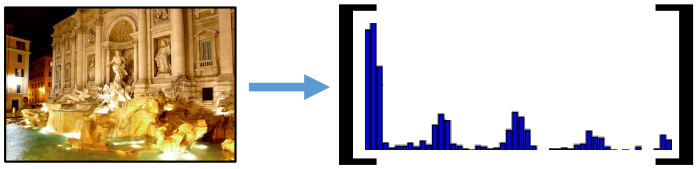
Image retrieval

URCV ETH zürich Microsoft



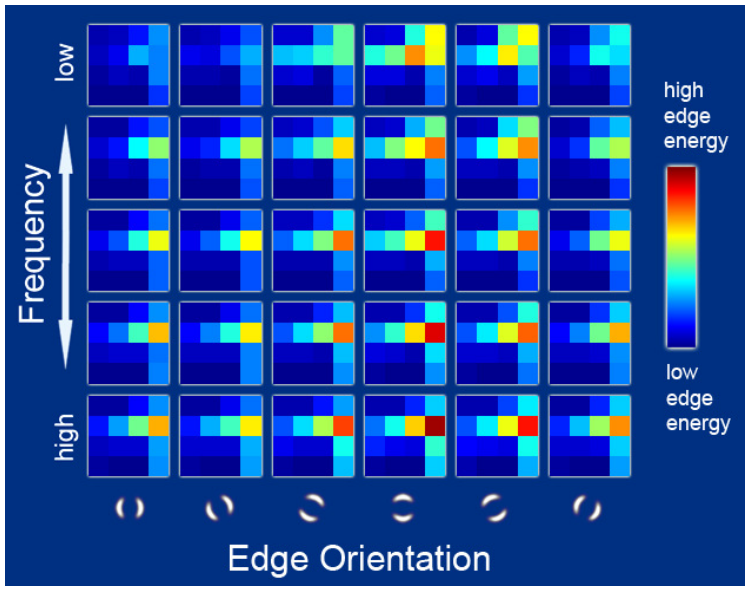
Descriptors

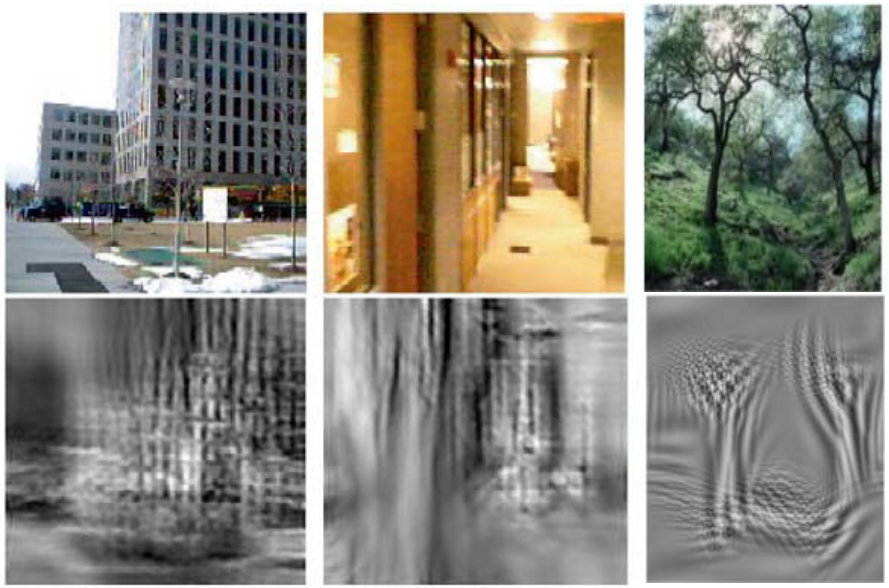
- Global image descriptor
  - Color histograms

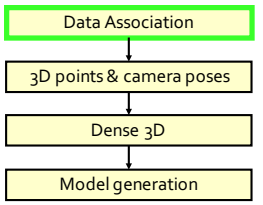


GIST-feature

- Several frequency bands and orientations for each image location
- Tiling of the image, for example 4x4, and at different resolutions
- Color histogram

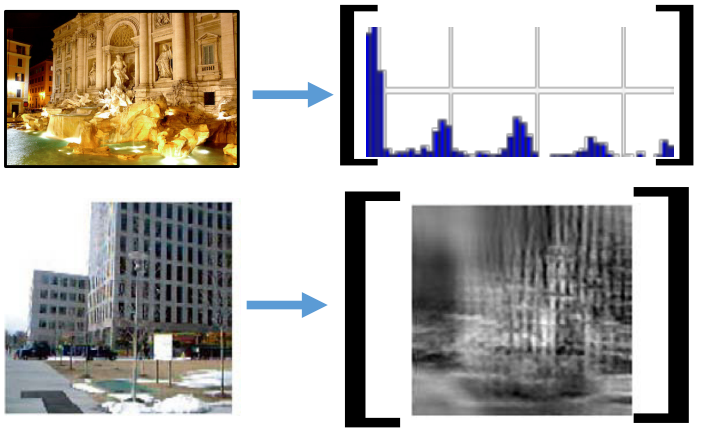


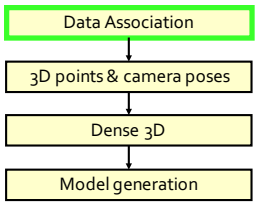




Descriptors

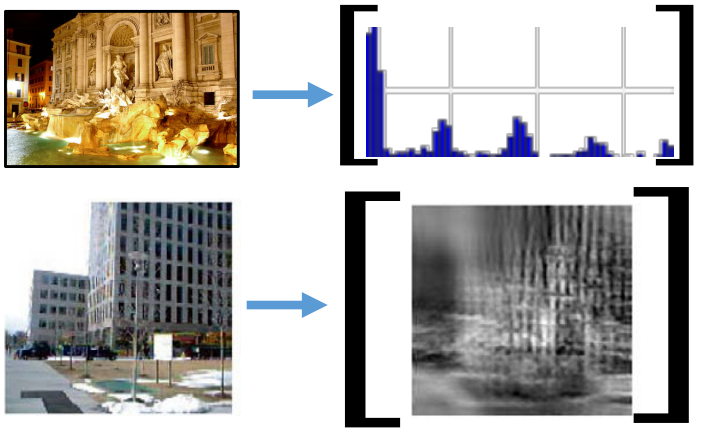
- Global image descriptor
  - Color histogram
  - GIST



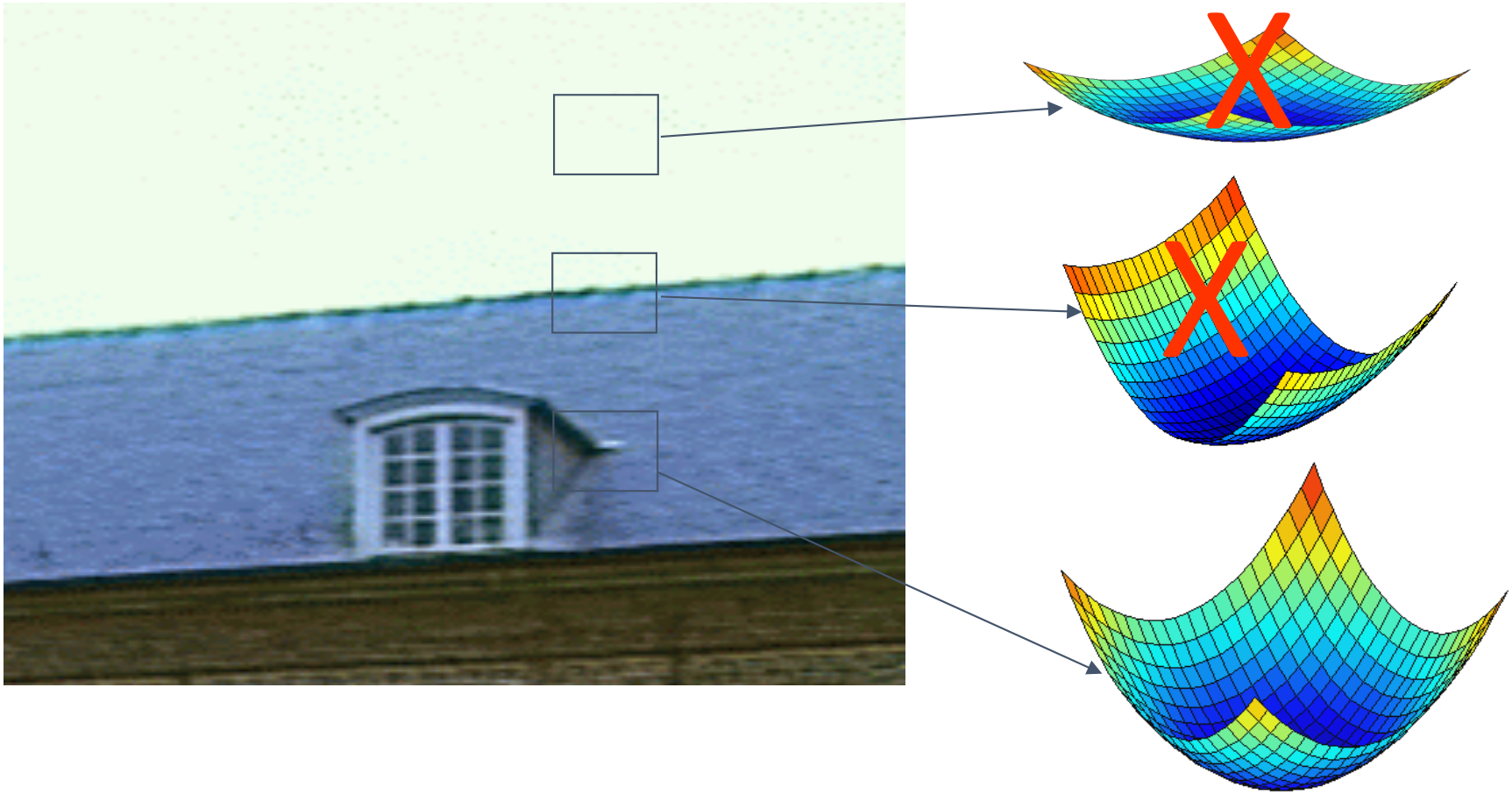


Descriptors

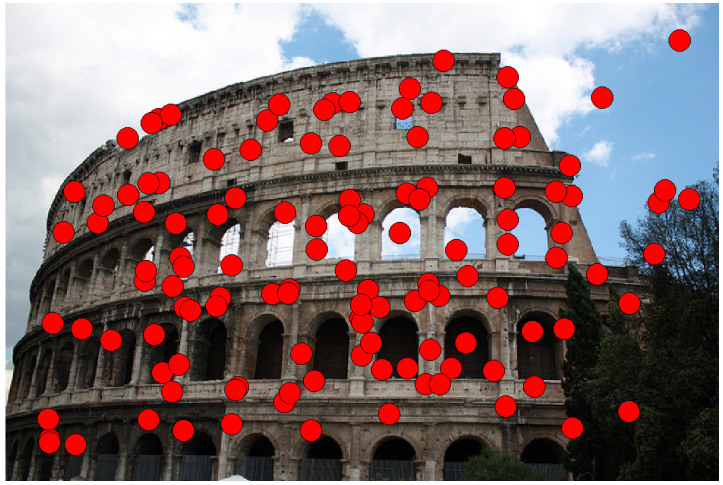
- Global image descriptor
  - Color histogram
  - GIST
- Local image descriptor

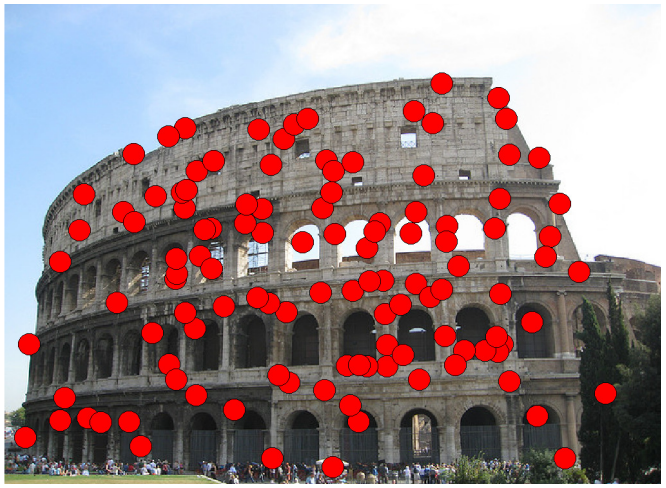


What are good features



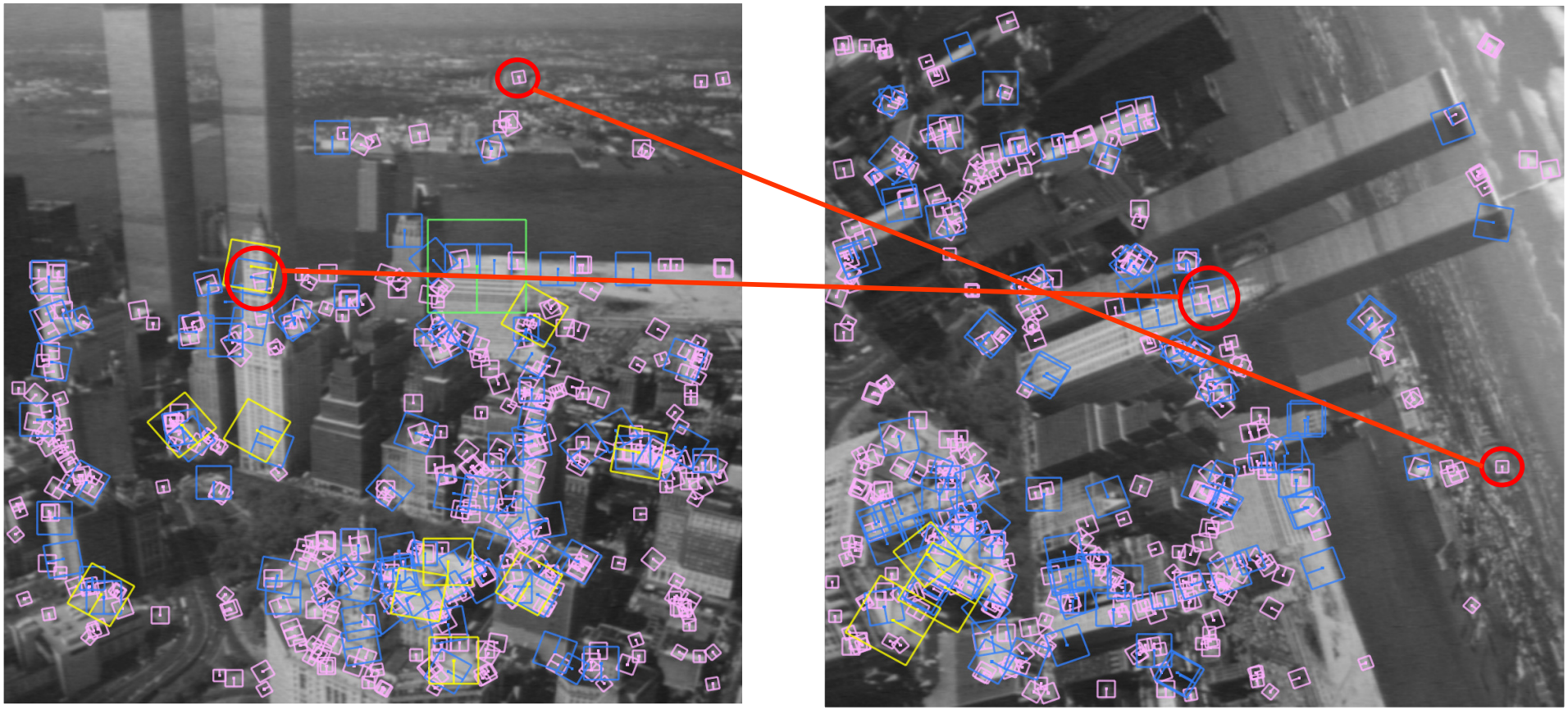
Characteristics of good features





- Repeatability
  - The same feature can be found in several images despite geometric and photometric transformations
- Saliency
  - Each feature is distinctive
- Compactness and efficiency
  - Many fewer features than image pixels
- Locality
  - A feature occupies a relatively small area of the image; robust to clutter and occlusion

SIFT-detector



Challenge: Detect features at:

1. different scales (sizes)
2. different orientations

SIFT-detector

- Scale and image-plane-rotation invariant feature descriptor [Lowe 2004]

  -Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters

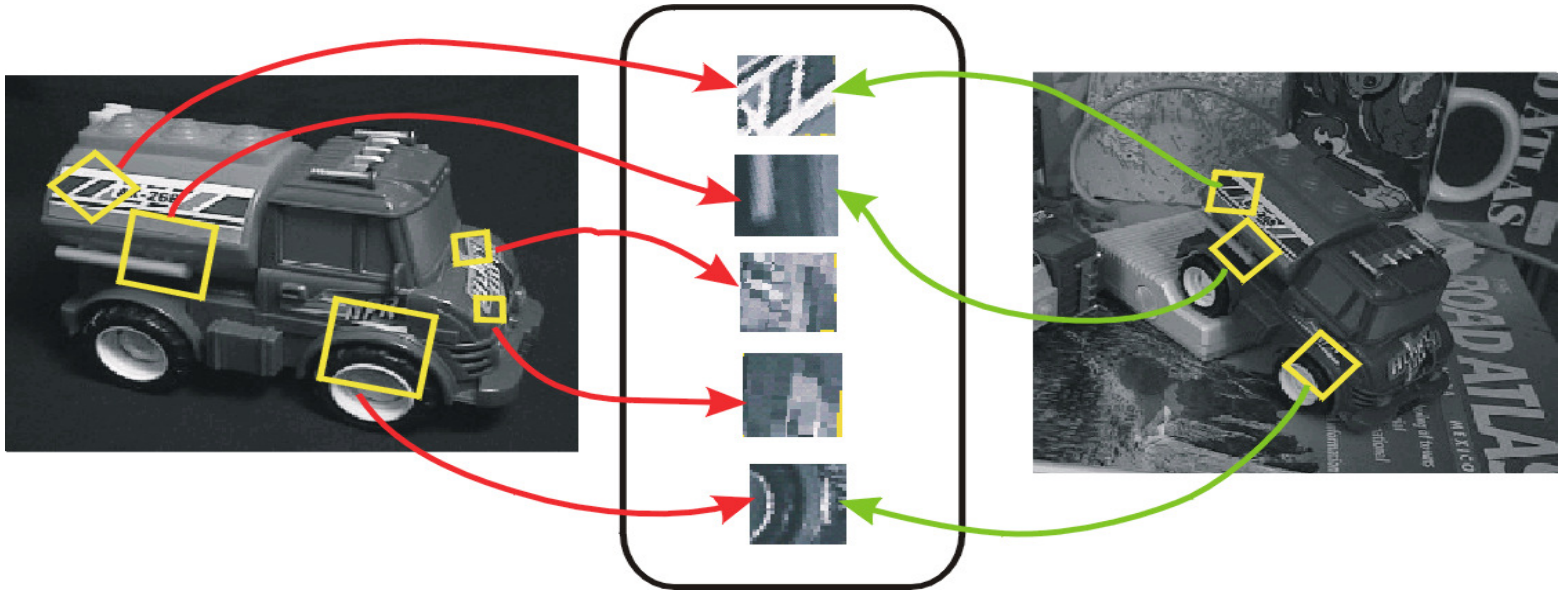
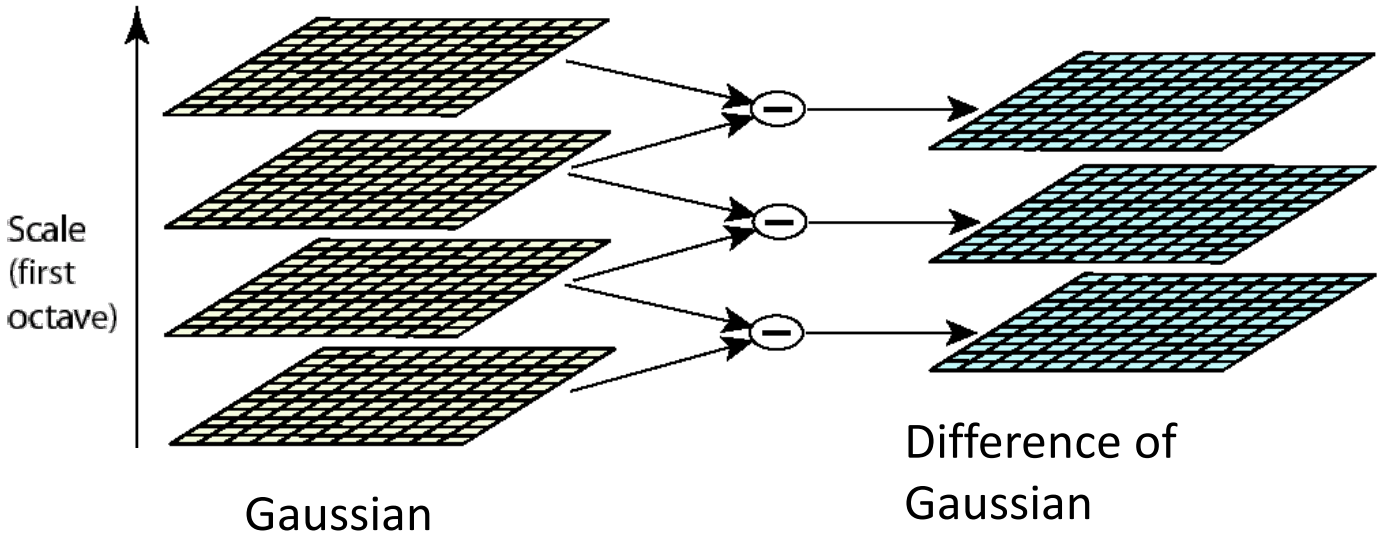


Image credit: D. Lowe

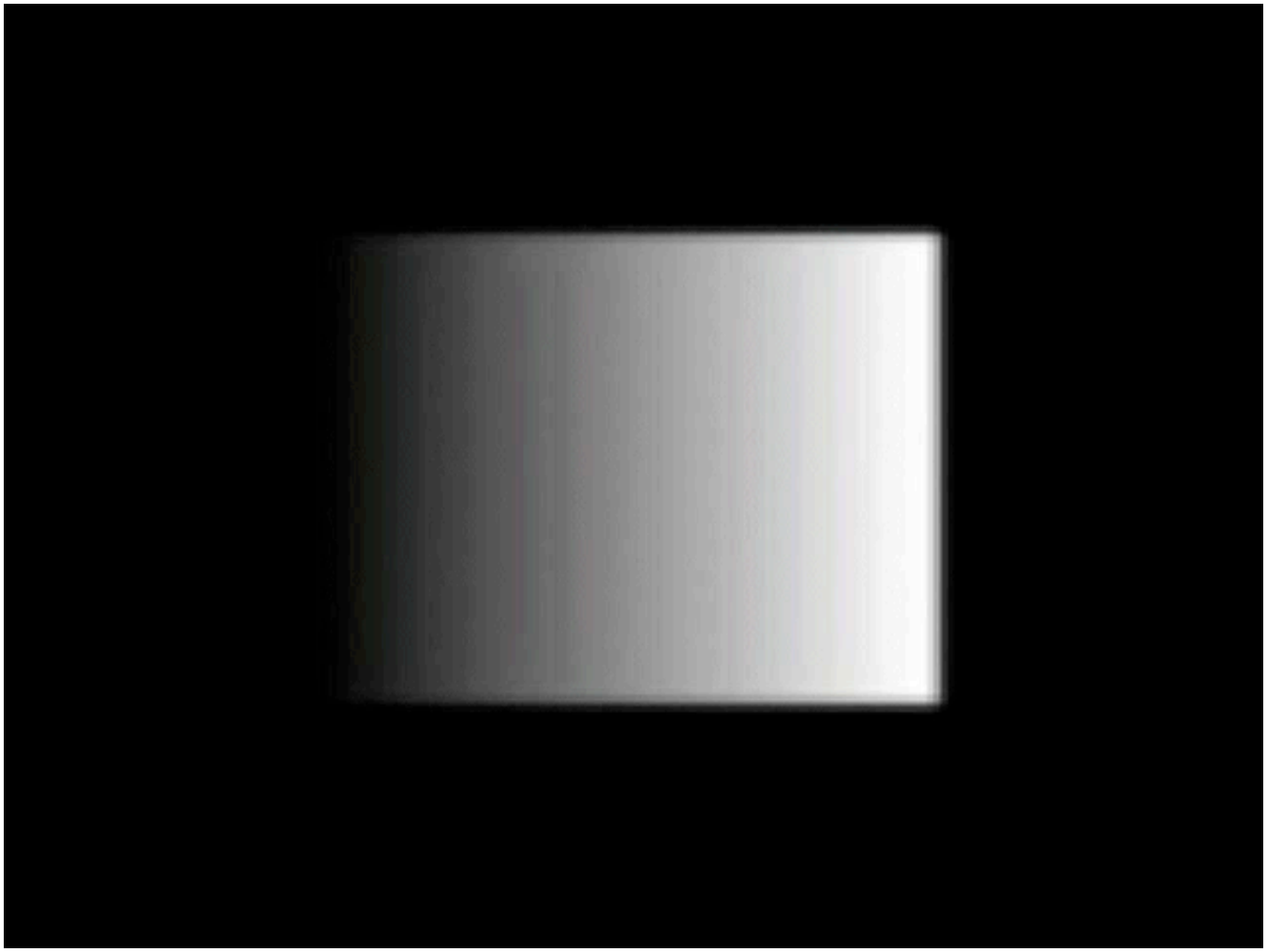
Difference of Gaussian for Scale Invariance



- Difference-of-Gaussian with constant ratio of scales is a close approximation to Lindeberg's scale-normalized Laplacian [Lindeberg 1998]

Image credit: D. Lowe

Difference of Gaussian for Scale Invariance



- Difference-of-Gaussian with constant ratio of scales is a close approximation to Lindeberg's scale-normalized Laplacian [Lindeberg 1998]

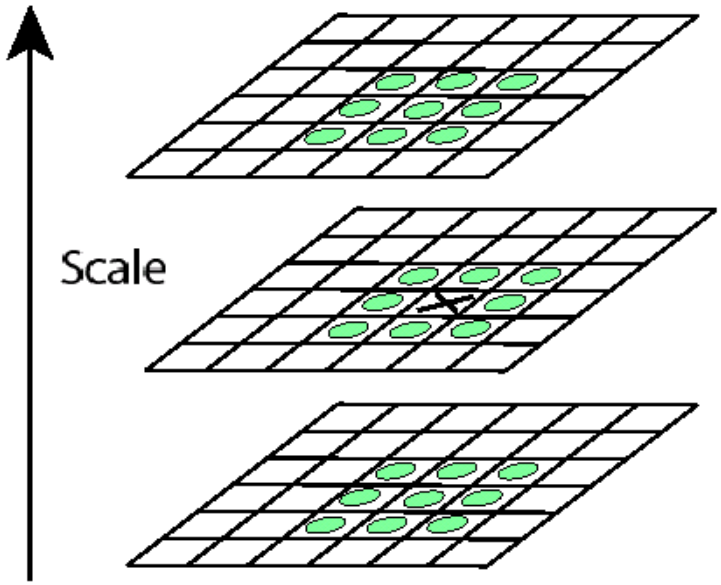
Key Point localization

- Detect maxima and minima of difference-of-Gaussian in scale space
- Fit a quadratic to surrounding values for sub-pixel and sub-scale interpolation (Brown & Lowe, 2002)
- Taylor expansion around point:

$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^\mathbf{T}\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}$$

- Offset of extremum (use finite differences for derivatives):

$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1}\frac{\partial D}{\partial \mathbf{x}}$$



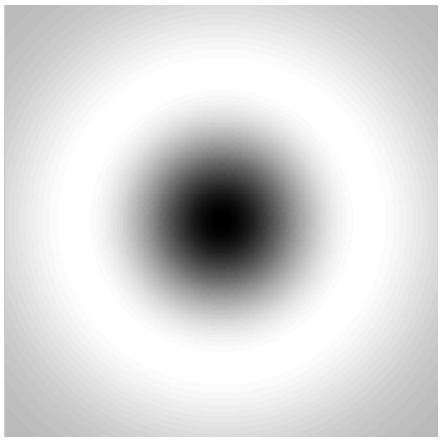
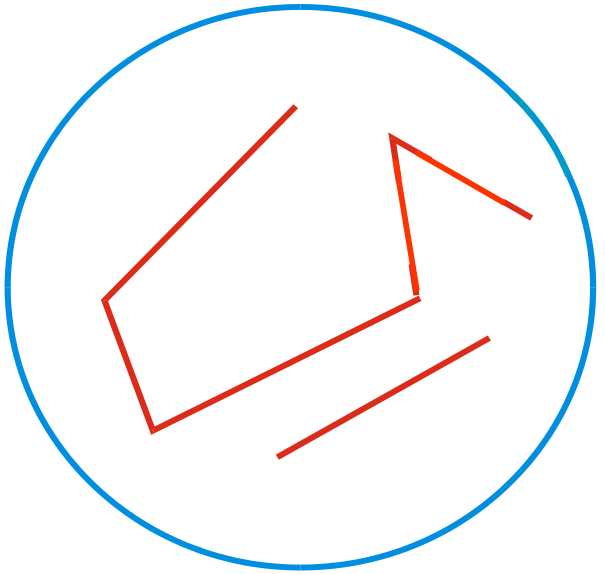


Image credit: D. Lowe

Orientation normalization

- Histogram of local gradient directions computed at selected scale

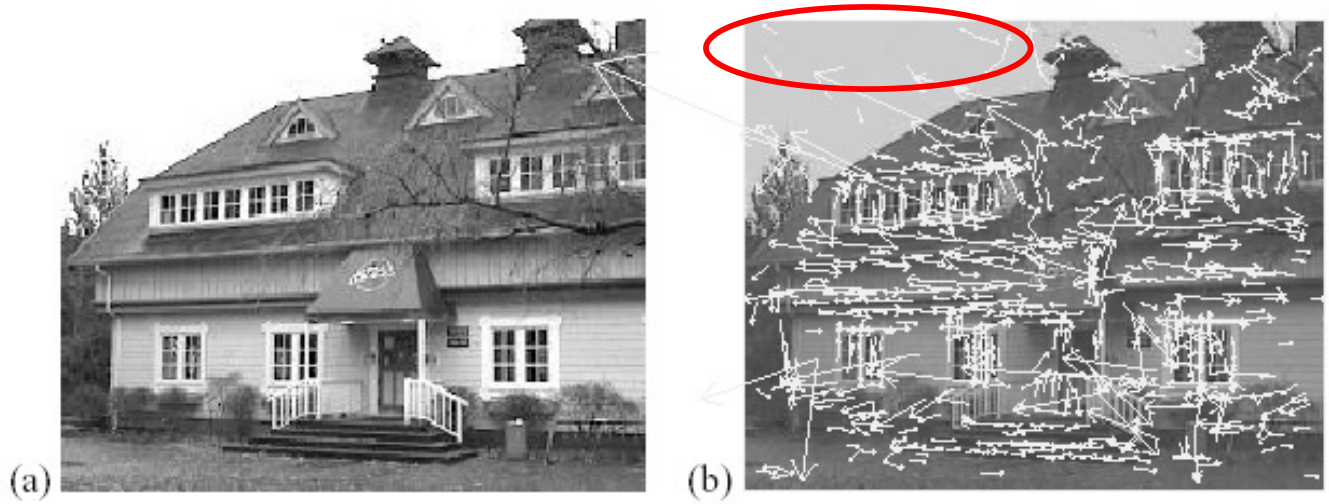


- Assign principal orientation at peak of smoothed histogram

- Each key specifies stable 2D coordinates (x, y, scale, orientation)

Example of Keypoint Detection

Threshold on value at DOG peak and on ratio of principle curvatures (Harris approach)



(a) 233x189 image
(b) 832 DOG extrema

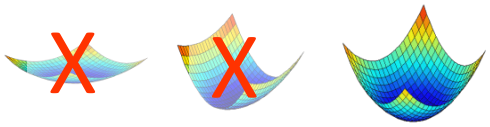


Image credit: D. Lowe

SIFT Descriptor Formation

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128 dimensions

example 2x2 histogram array

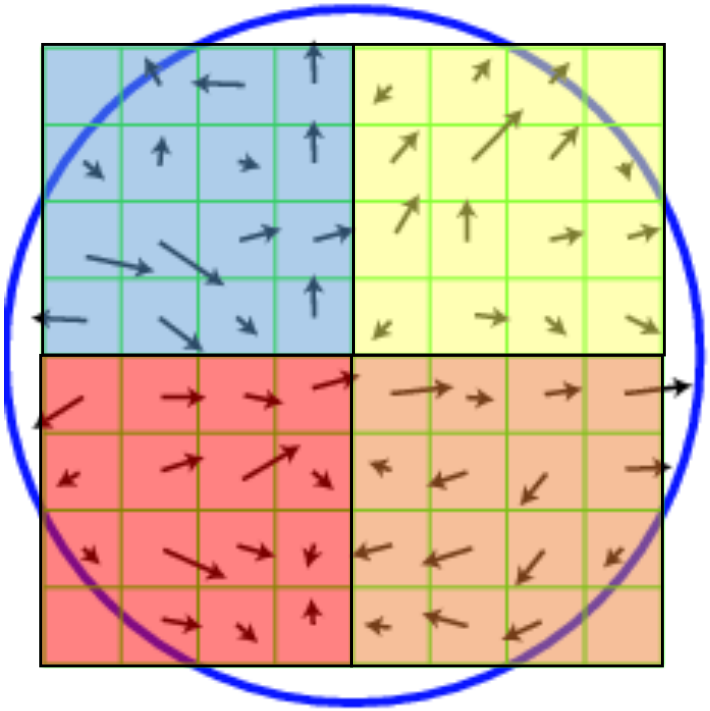
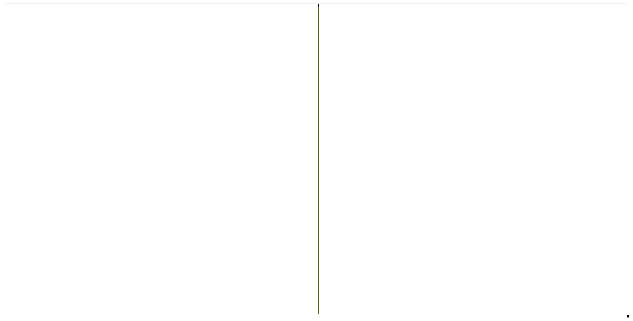


Image gradients



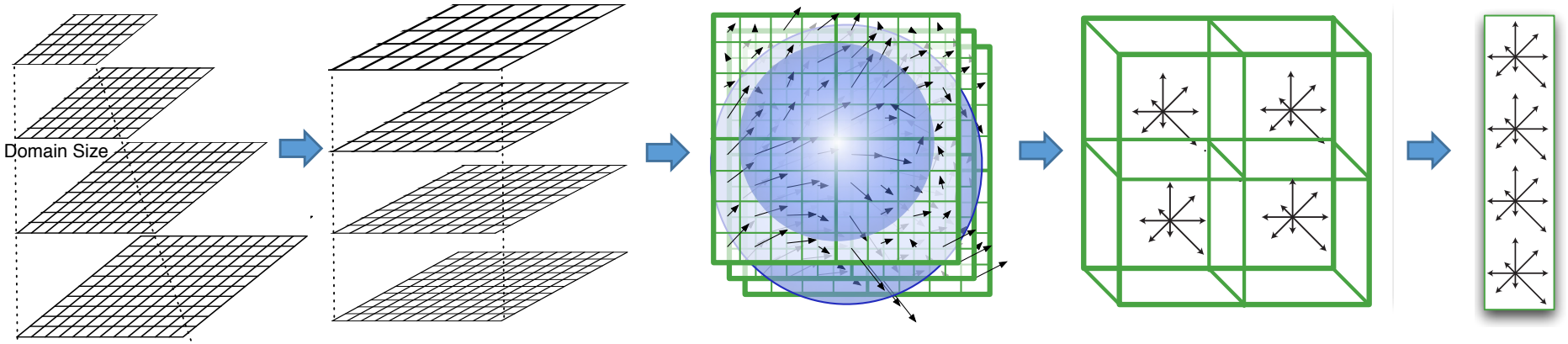
Keypoint descriptor

SIFT Feature Detector

Domain Size Pooling (DSP)-SIFT

[Dong and Soatto, CVPR 2015]

- Idea: Use multiple scales to build descriptor



15 rescaled patches with scales $\in \left[\frac{\sigma}{6}, \frac{4\sigma}{3}\right]$

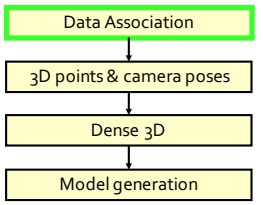
Compute gradients per scale

Compute histograms per scale

Accumulate histograms over scales

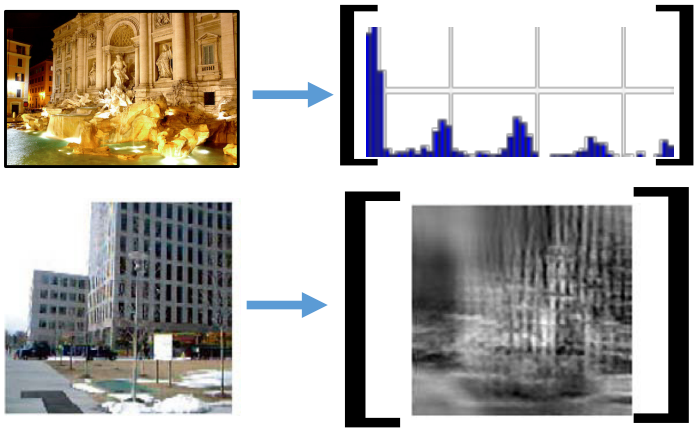
- Best performing local descriptor for SfM [Schönberger CVPR 2017]

Image credit: Dongç



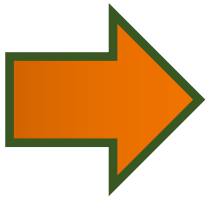
Descriptors

- Global image descriptor
  - Color histogram
  - GIST
- Local image descriptor
  - SIFT, DSP-SIFT



Goal

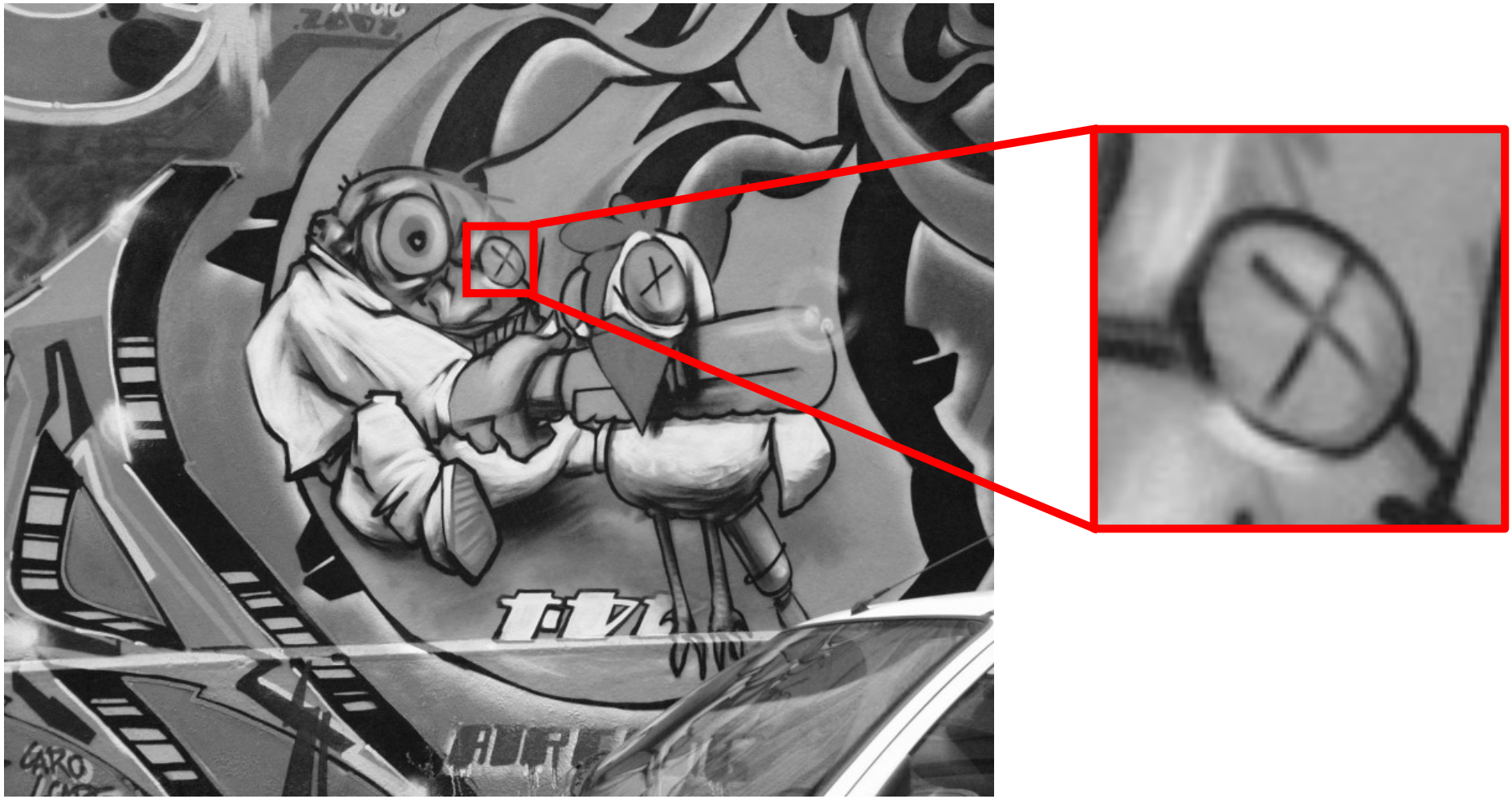
Image Patch



[0, 0, 1, 0, 1, 1, 0, 1, …]

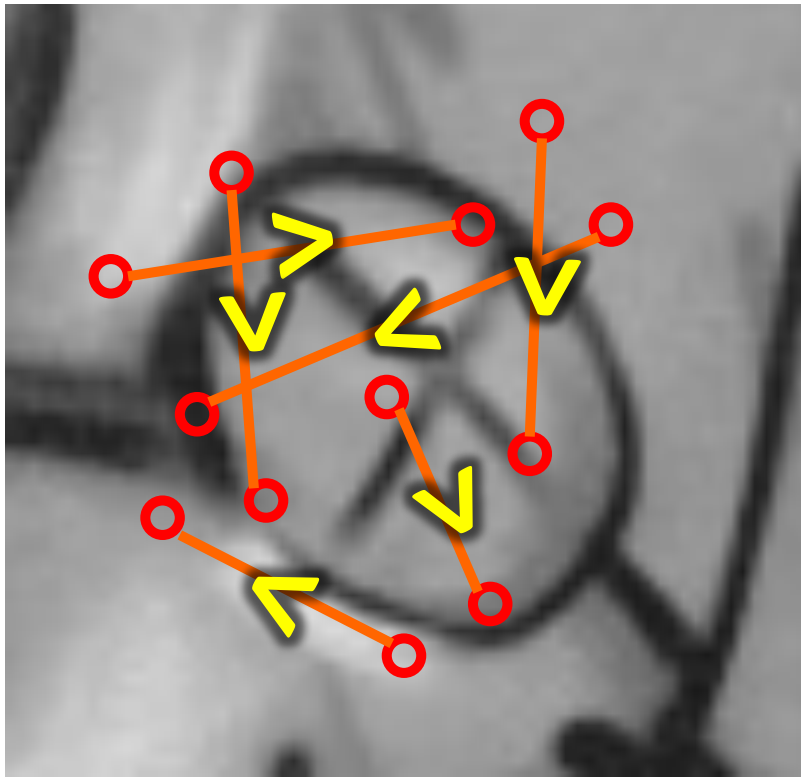
Binary Descriptor

Feature Description



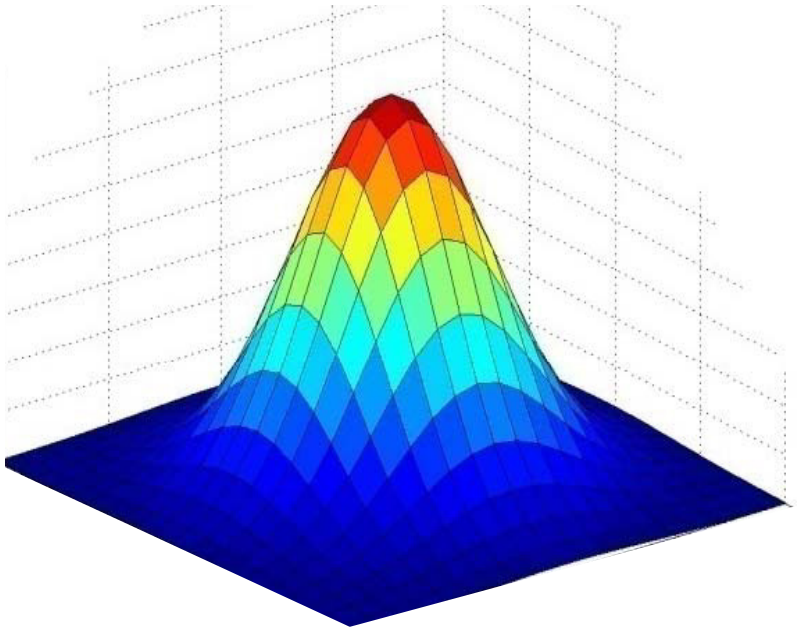
BRIEF: Method

[Calonder et al. ECCV 2010]



Descriptor: **0 1 1 0 1 0 …**

BRIEF: Sampling



Endpoints from
2D Gaussian

BRIEF: Descriptor

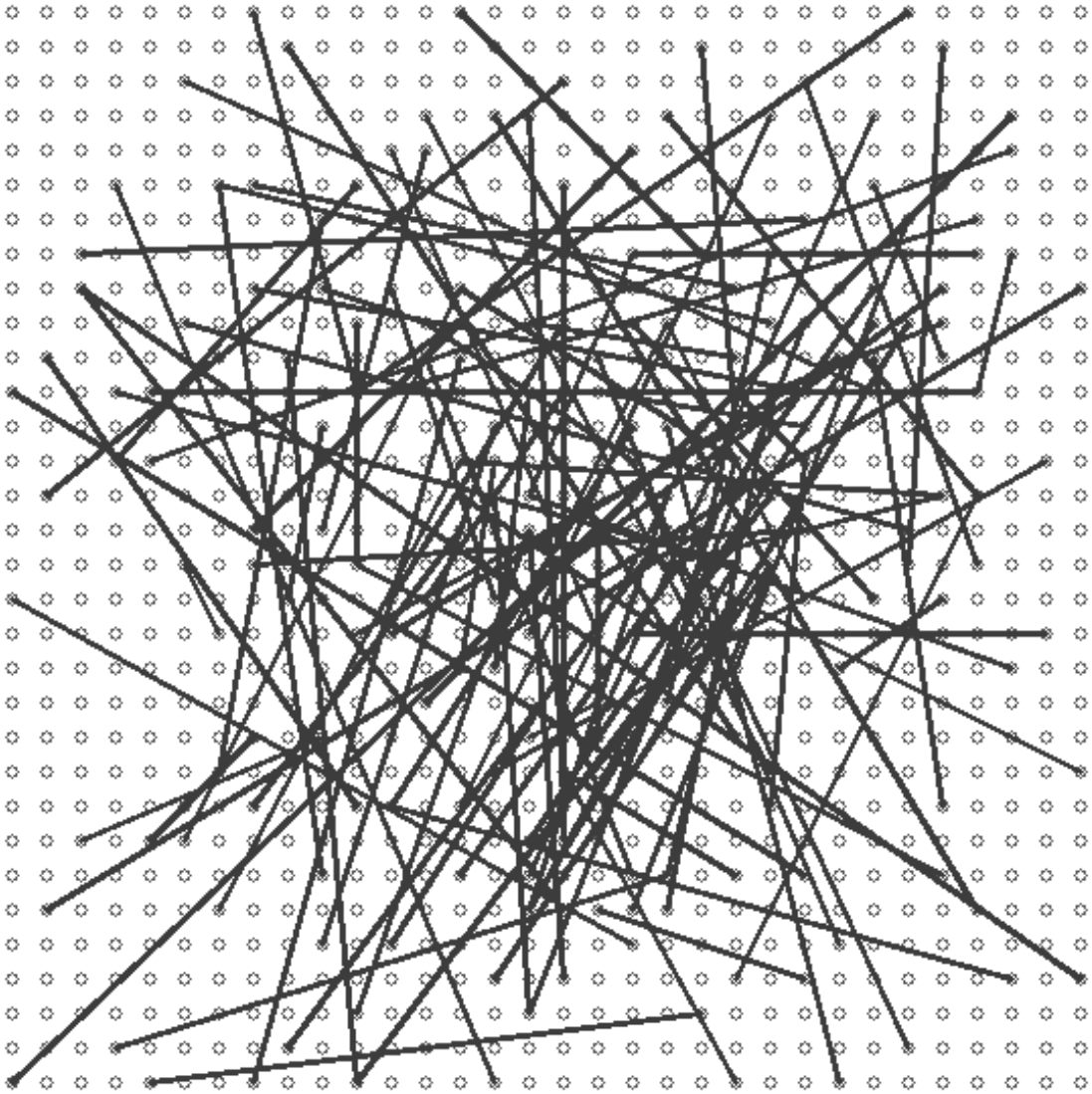


Image credit: Calonder

BRIEF: Descriptor

- 128, 256, or 512 bits
  - 16, 32, or 64 bytes
- Hamming distance matching

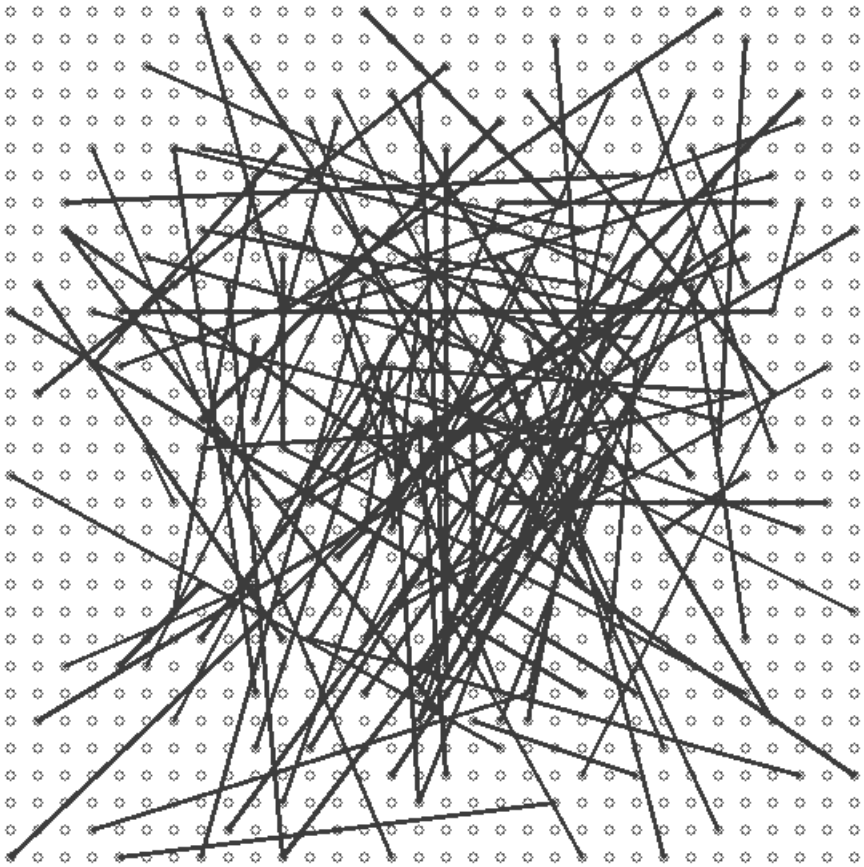


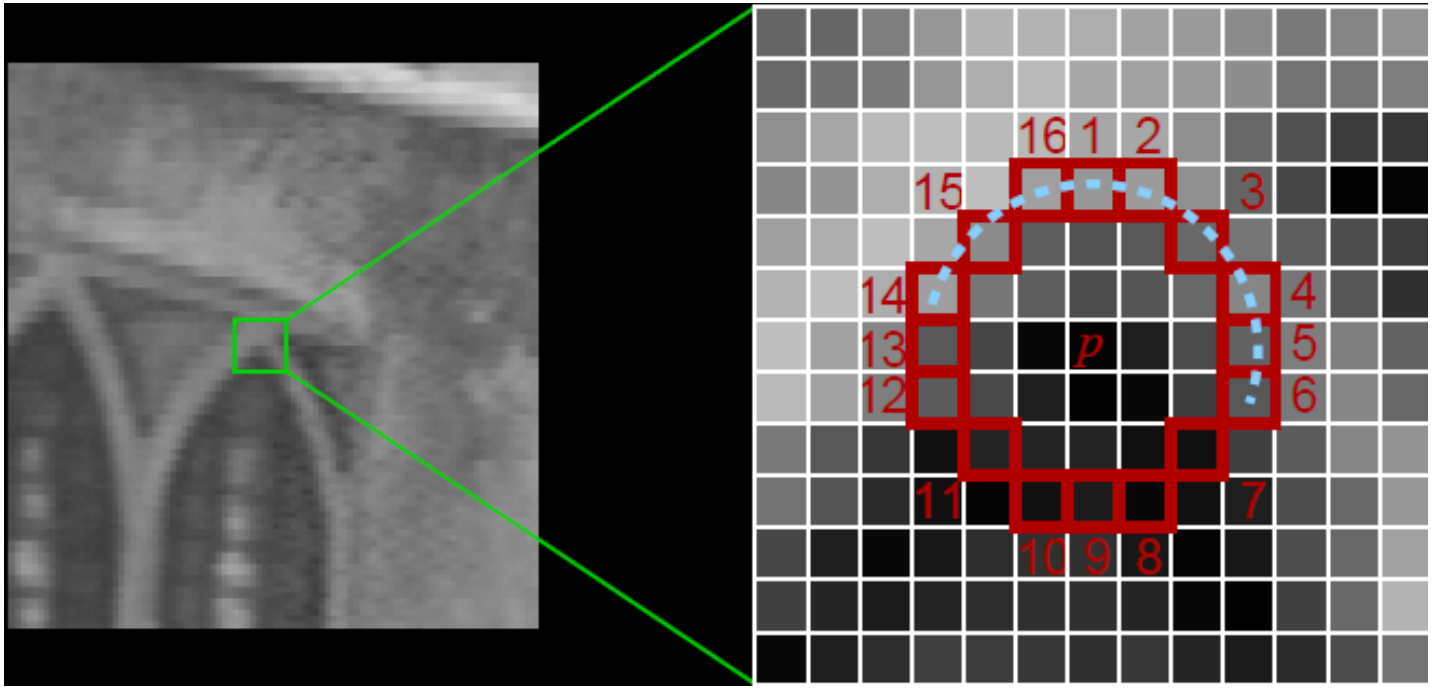
Image credit: Calonder

BRIEF: Summary

- Pros
  - Highly efficient
- Cons
  - No scale invariance
  - No rotation invariance
  - Sensitive to noise

ORB: Fast Corner Detector

[Rublee et at. ICCV 2011]



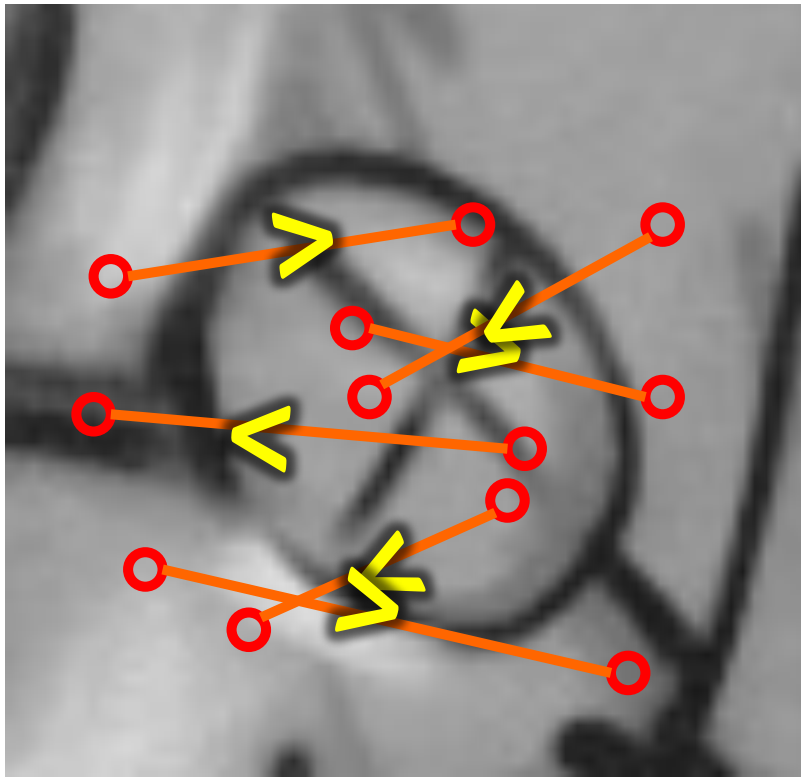
- Continuous arc of pixels all much
  - brighter than center pixel p (brighter than p+threshold)

  or
  - darker than center pixel p (darker than p-threshold)
- ≥ 12 pixel brighter or darker
- Rapid rejection by testing pixel 1,9,5, and 13
- Non-maxima suppression

adapted from Ed Rosten

ORB: Method

[Rublee et at. ICCV 2011]

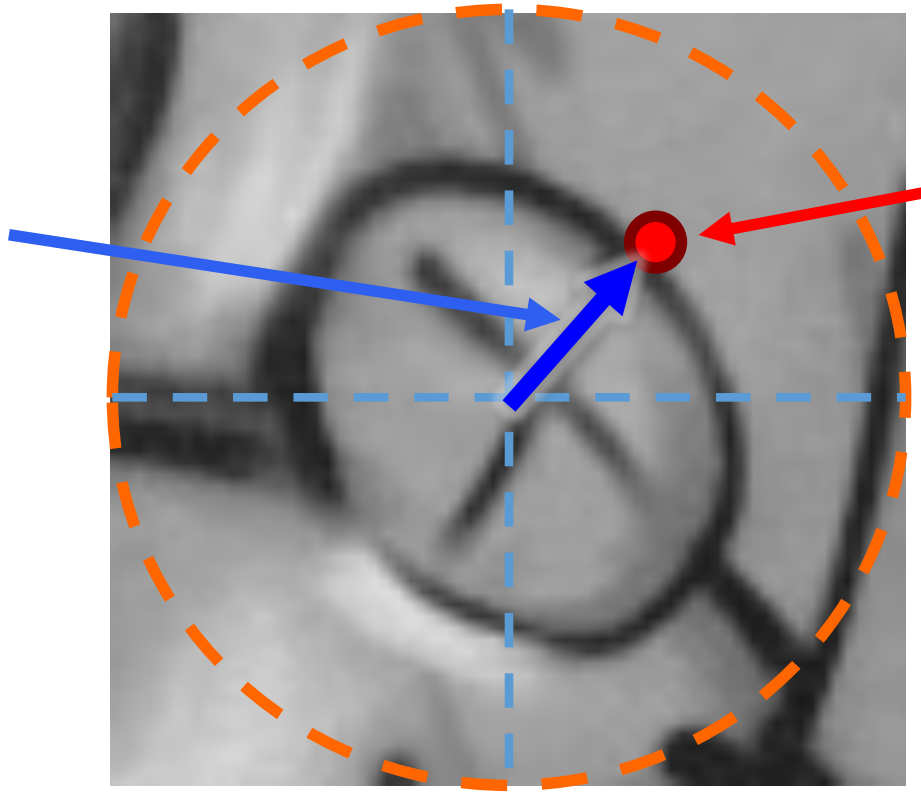


Descriptor: **0 1 1 0 1 0 …**

ORB: Rotation Invariance

moment of patch: $m_{pq} = \sum_{x,y} x^p y^q I(x,y)$

Feature Direction



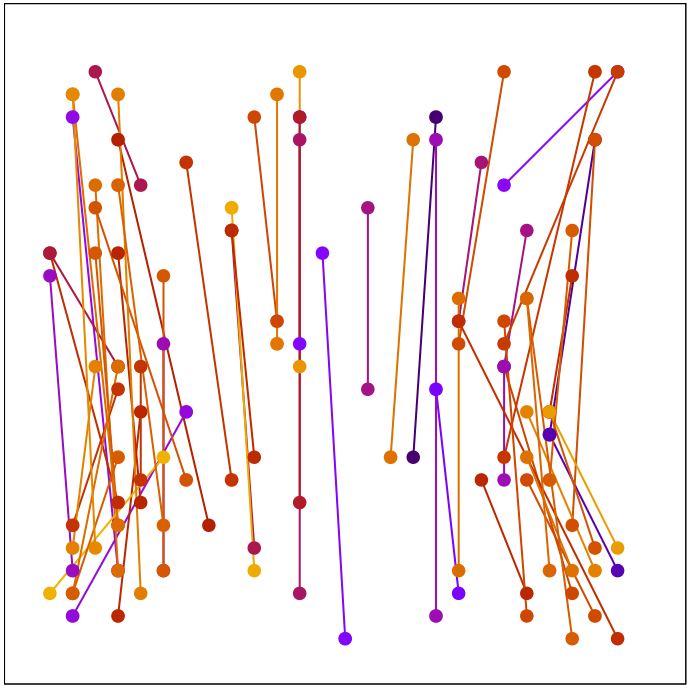
Intensity Centroid

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right)$$

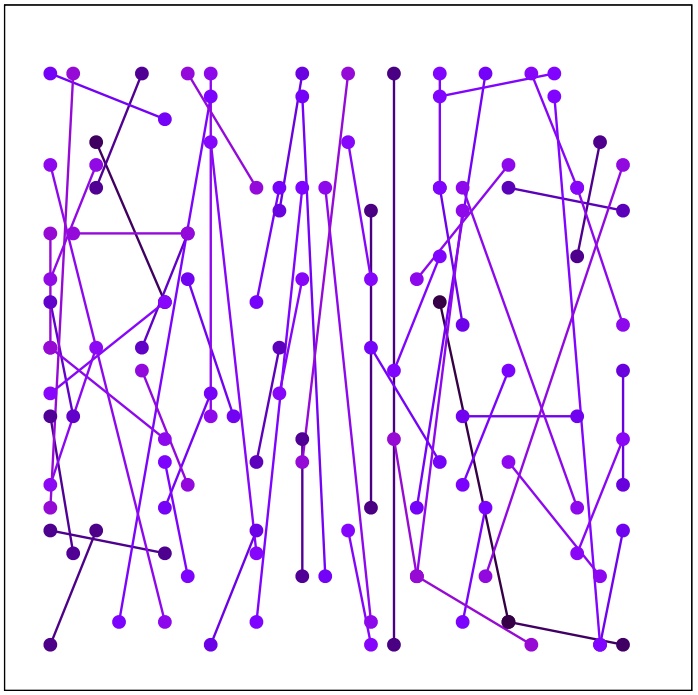
- choose orientation corrected gradients for descriptor generation

ORB: Descriptor

Candidate Arrangement



Learned Arrangement



Low Endpoint Correlation **High**

Image credit: Rublee

ORB: Summary

- Pros
  - Efficient
  - Rotation invariance
- Cons
  - No scale invariance
  - Sensitive to noise

slide: J. Heinly

Data Association

Data Association

images, 2D features

Image correspondence

Image retrieval

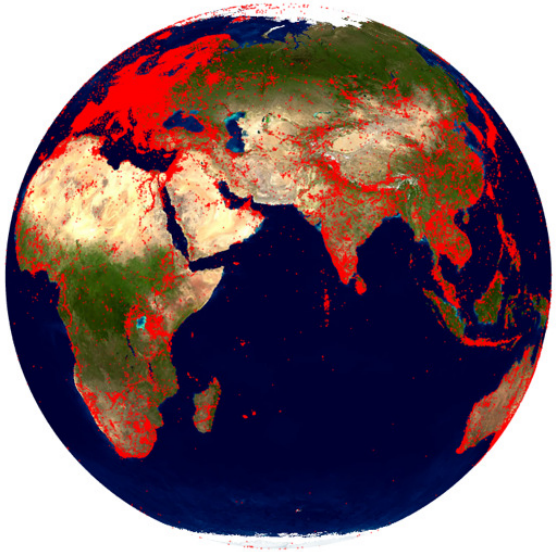
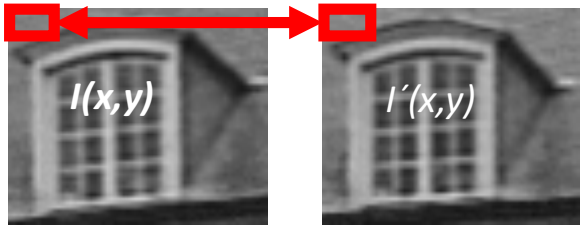
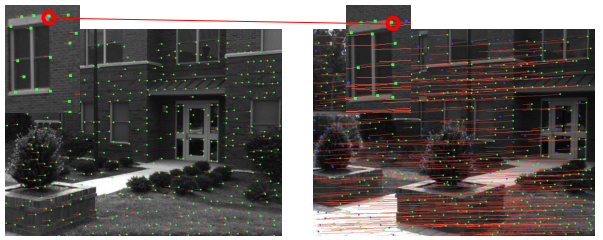


Image Correspondence

- There are two principled ways for finding correspondences:



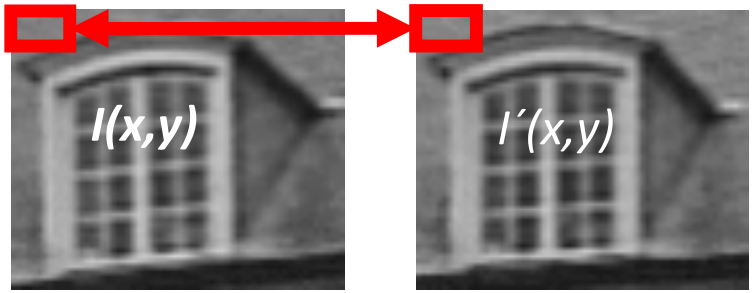
1. Matching
   - Independent detection of features in each frame
   - Correspondence search over detected features
   + Extends to large transformations between images
   − High error rate for correspondences



2. Tracking
   - Detect features in one frame
   - Retrieve same features in the next frame by searching for equivalent feature (tracking)
   + Very precise correspondences
   + High percentage of correct correspondences
   − Only works for small changes in between frames

Comparing image regions

Compare intensities pixel-by-pixel



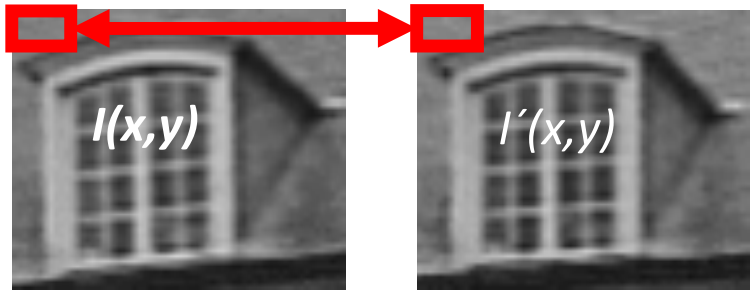
Dissimilarity measures

Sum of Square Differences

$$SSD = \sum_{(x,y)\in w} \sum \left(I'(x,y) - I(x,y) \right)^2$$

Comparing image regions

Compare intensities pixel-by-pixel



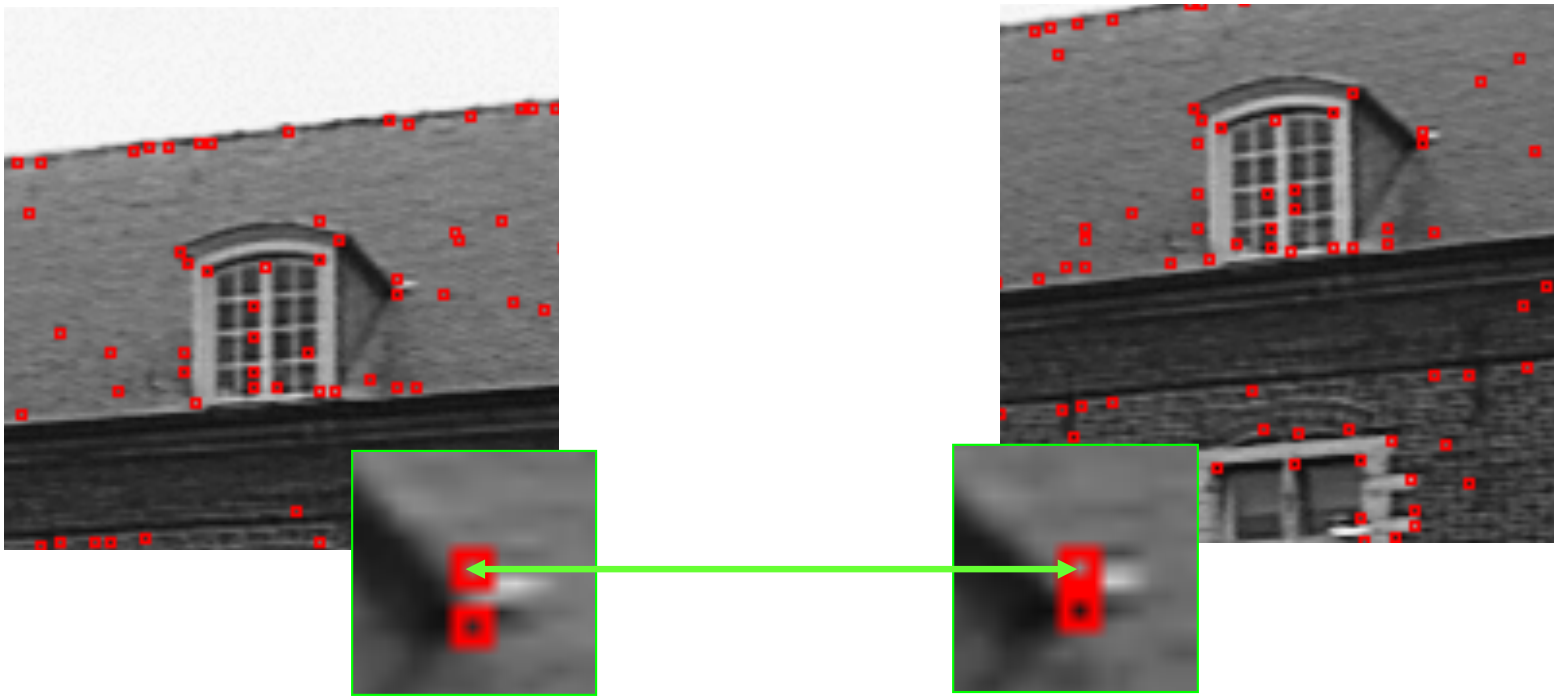
Similarity measures

Zero-mean Normalized Cross Correlation

$$NCC = \frac{\sum_{(x,y)\in w} \sum \left(I'(x,y) - \bar{I}'\right)^2 \left(I(x,y) - \bar{I}\right)^2}{\sigma_{I'(w)} \sigma_{I(w)}}$$

Simple Matching

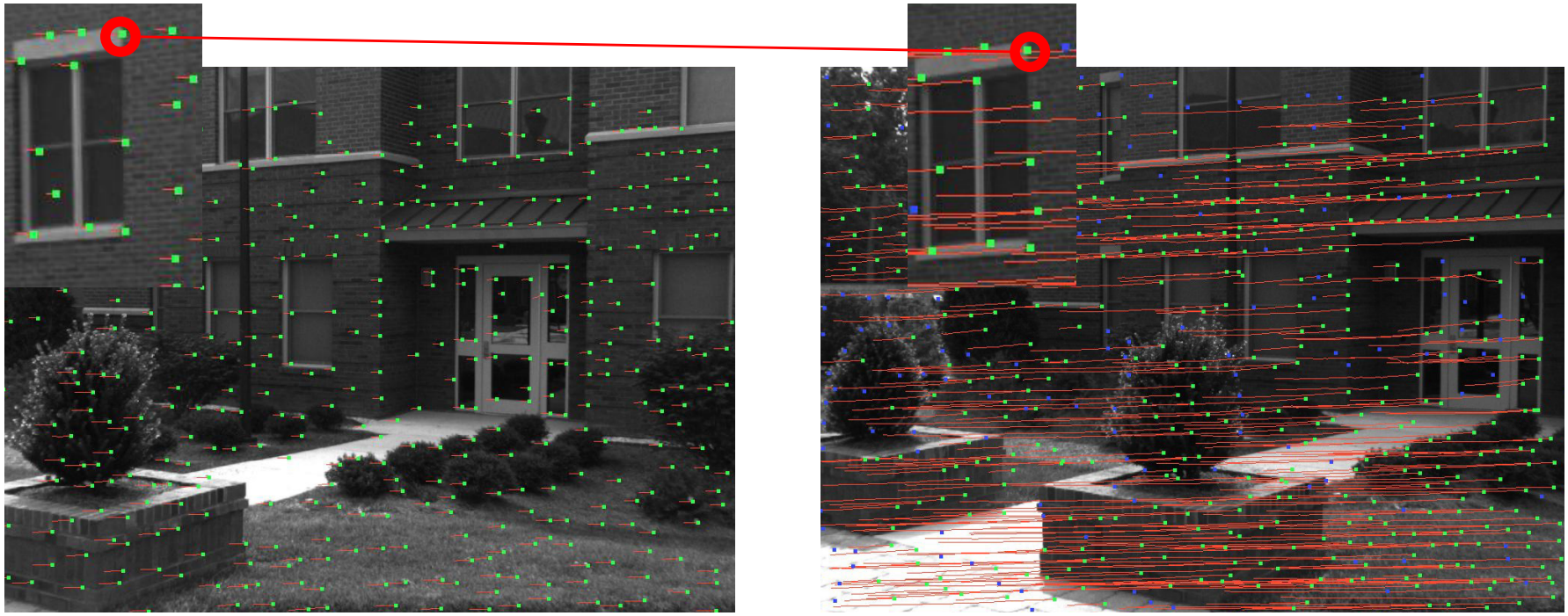
- For each corner (feature) in image 1
  - find the corner in image 2 that is most similar
  - use SSD or NCC
  - Perform reverse match
- Keep mutual best matches



What transformations does this work for?

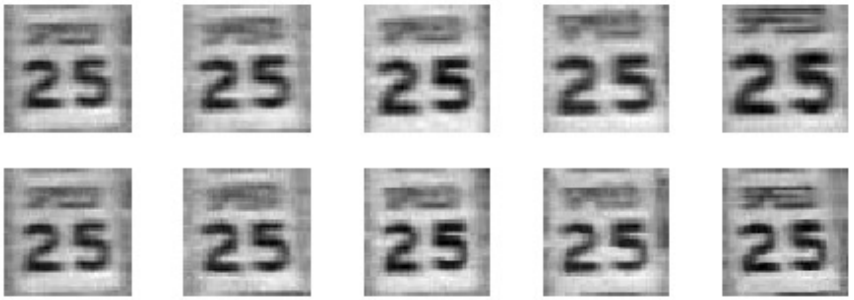
Feature Tracking

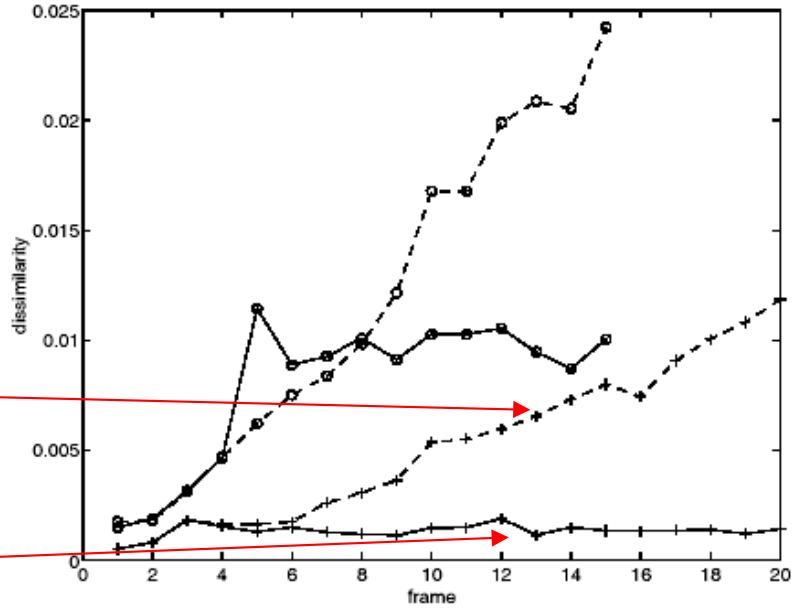
- Establish correspondences between identical salient points multiple images



Example

Figure 1: Three frame details from Woody Allen's *Manhattan*. The details are from the 1st, 11th, and 21st frames of a subsequence from the movie.





Simple displacement is sufficient between consecutive frames, but not to compare to reference template

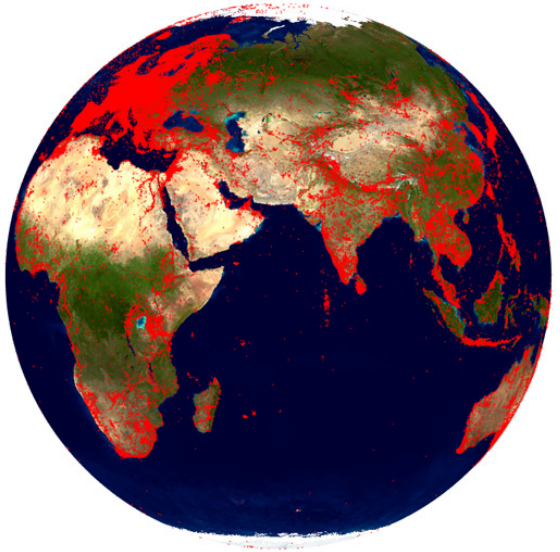
Data Association

Data Association

images, 2D features

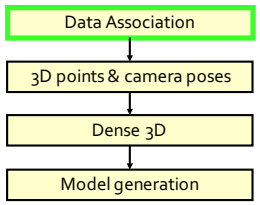
Image correspondence

Image retrieval

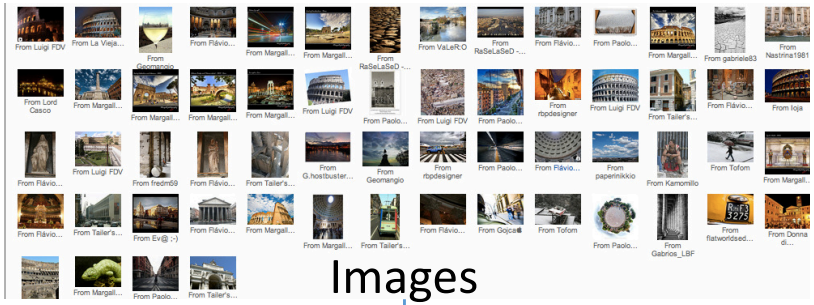


URCV ETH zürich Microsoft

Data Association



New image



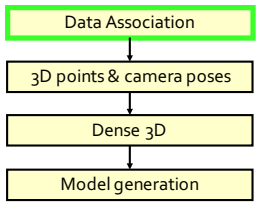
Images

Features

Index

Database

URCV ETH zürich Microsoft



Data Indexing

- Indexing through Vocabulary Tree [Nister CVPR 2006]

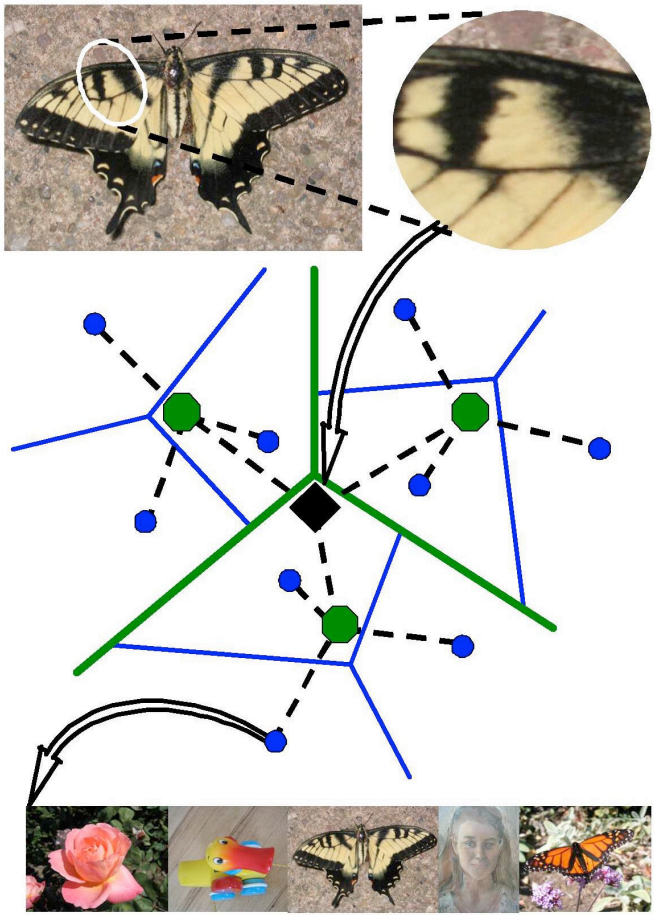
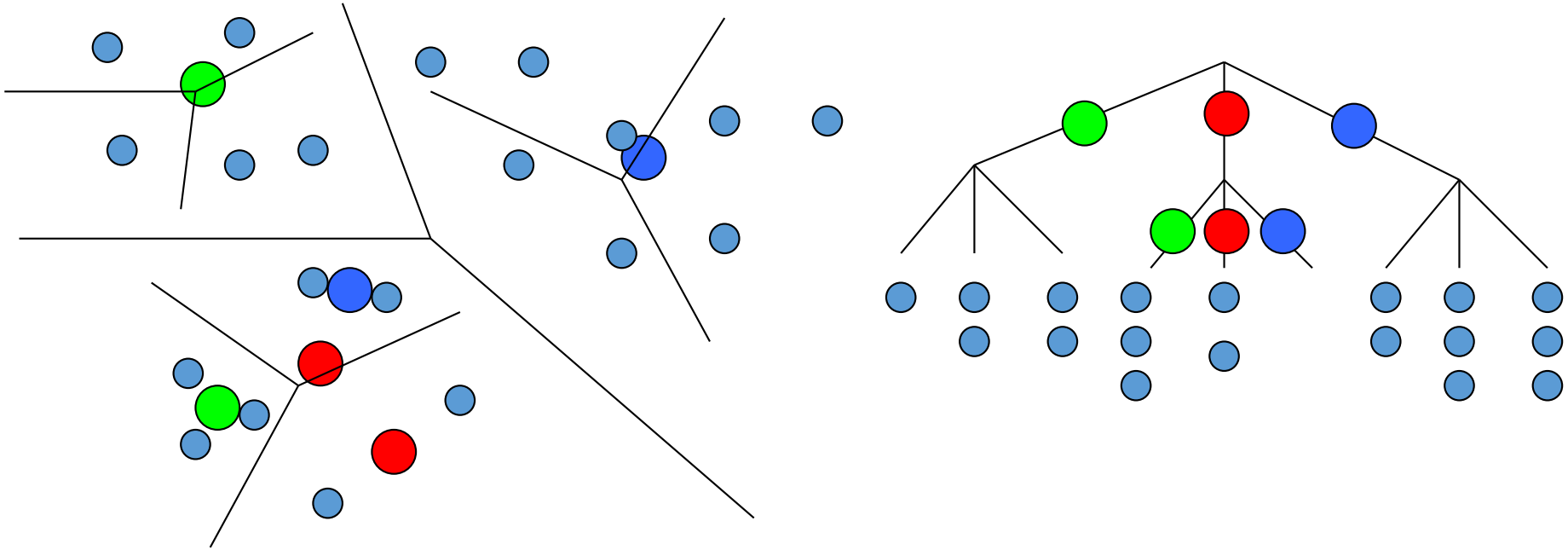


Image credit: D. Nister

What is a Vocabulary Tree?

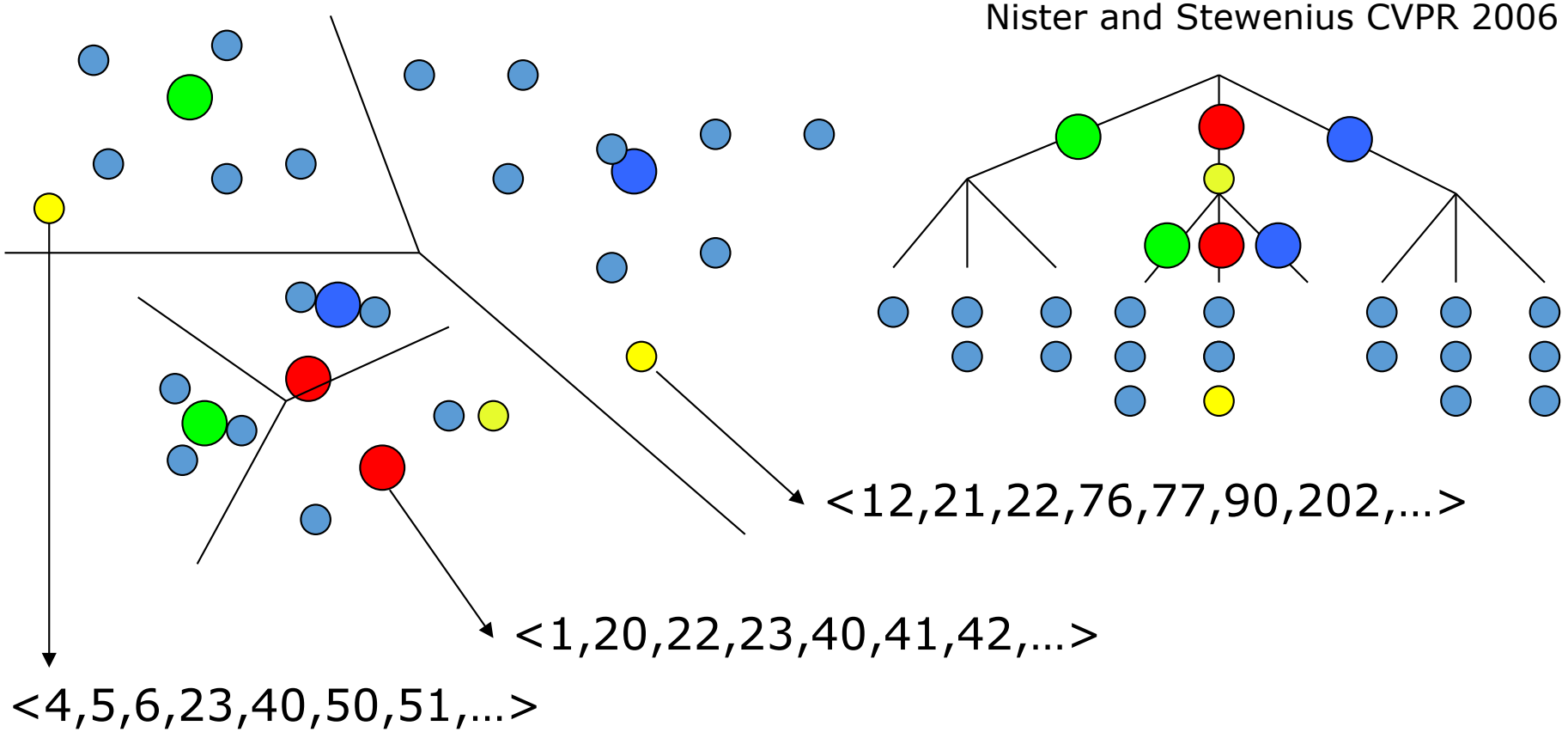
Nister and Stewenius CVPR 2006



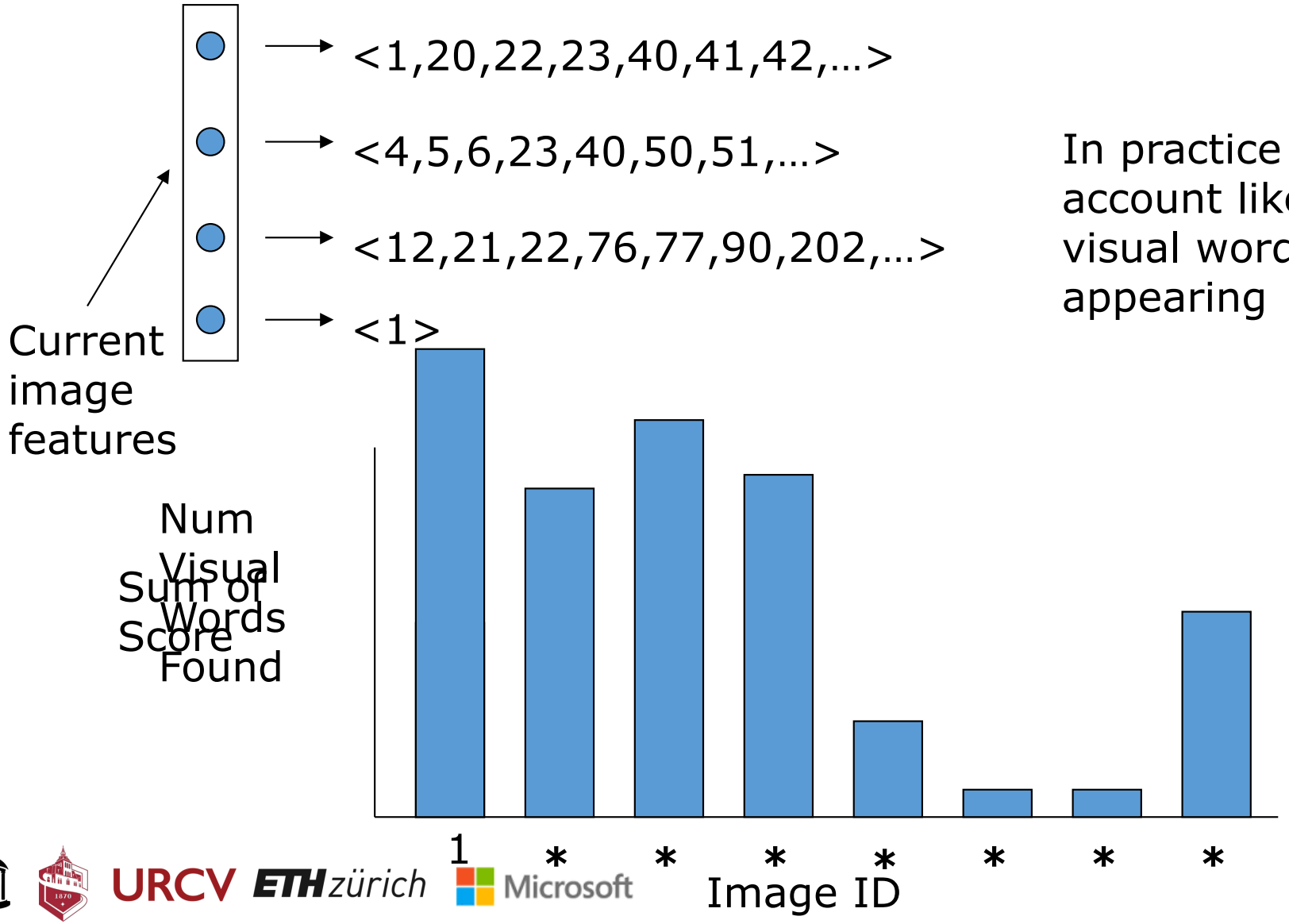
- Multiple rounds of K-Means to compute decision tree (offline)
- Fill and query tree online

Quantizing a SIFT Descriptor

Nister and Stewenius CVPR 2006



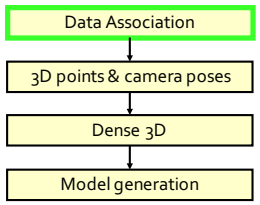
Scoring Images



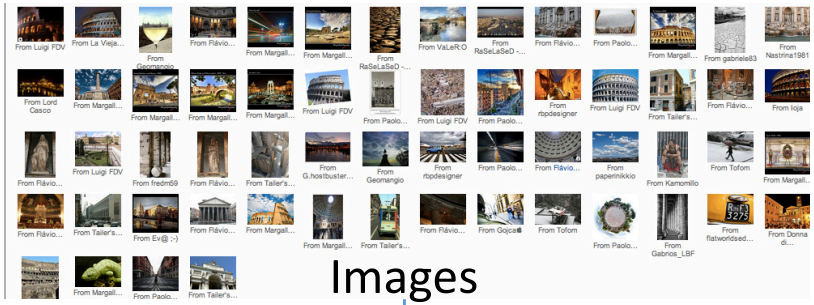
In practice take into account likelyhood of visual word appearing

Nister and Stewenius CVPR 2006

Data Association



New image



Images

Features

Index

Database

Candidates

URCV ETH zürich Microsoft