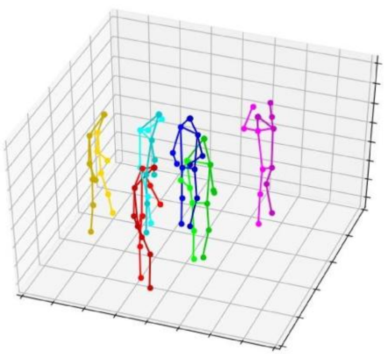
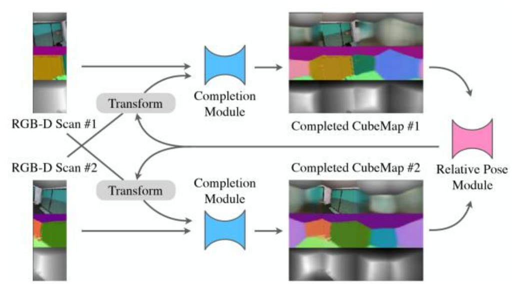
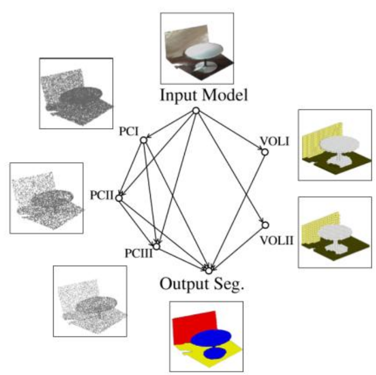
CS376 Computer Vision
Lecture 16: Two-View Stereo





Qixing Huang

March 27th 2019

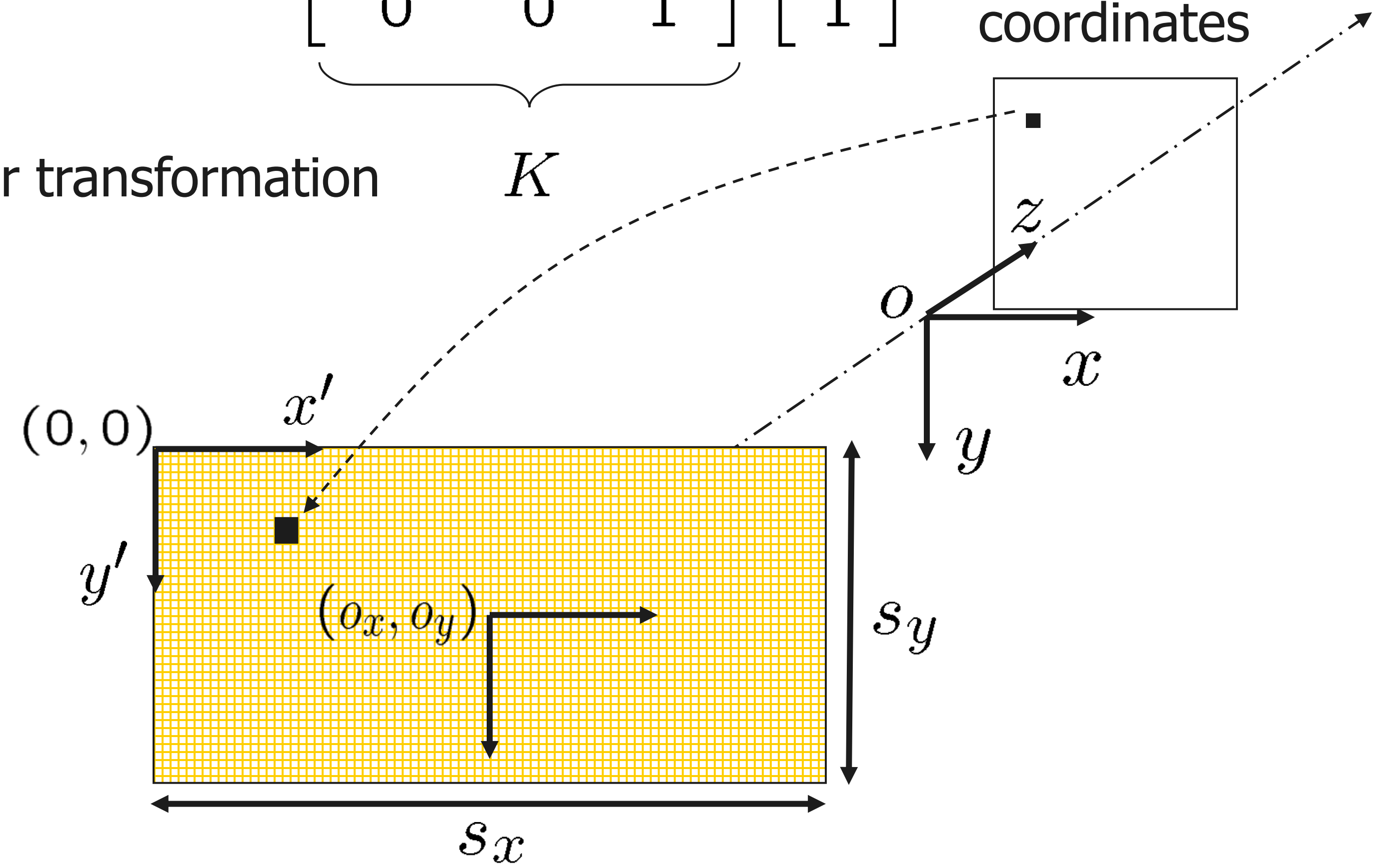


Camera Calibration

Uncalibrated Camera – Intrinsic Parameters are unknown

$$\boldsymbol{x}' = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = K\boldsymbol{x} = \underbrace{\begin{bmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Linear transformation



Calibration with a Rig

Uncalibrated Camera Using Homogeneous Coordinates

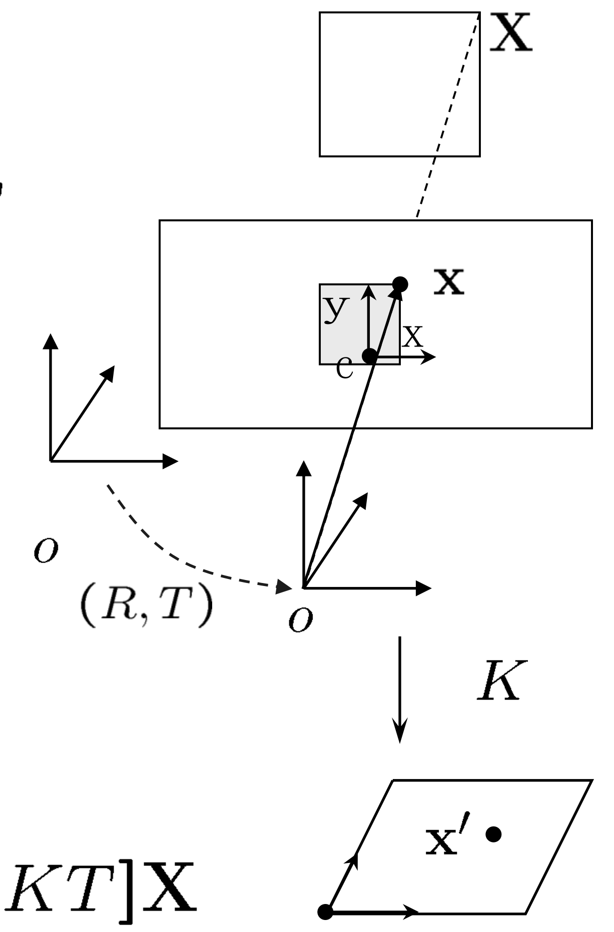
$$\mathbf{X} = [X, Y, Z, W]^T \in \mathbb{R}^4, \quad (W = 1)$$

Last Lecture:

- Image plane coordinates $\mathbf{x} = [x, y, 1]^T$
- Camera extrinsic parameters $g = (R, T)$
- Perspective projection $\lambda \mathbf{x} = [R, T]\mathbf{X}$

This Lecture:

- Pixel coordinates $\mathbf{x}' = K\mathbf{x}$
-
- Projection matrix $\lambda \mathbf{x}' = \Pi \mathbf{X} = [KR, KT]\mathbf{X}$



Calibration with a Rig

Use the fact that both 3-D and 2-D coordinates of feature points on a pre-fabricated object (e.g., a cube) are known.

Calibration with a Rig

- Given 3-D coordinates on known object $\mathbf{X}$

$$\lambda \mathbf{x}' = [KR, KT]\mathbf{X} \quad \Longrightarrow \quad \lambda \mathbf{x}' = \Pi \mathbf{X}$$

$$\lambda \begin{bmatrix} x^i \\ y^i \\ 1 \end{bmatrix} = \begin{bmatrix} \pi_1^T \\ \pi_2^T \\ \pi_3^T \end{bmatrix} \begin{bmatrix} X^i \\ Y^i \\ Z^i \\ 1 \end{bmatrix}$$

- Eliminate unknown scales

$$\begin{aligned} x^i(\pi_3^T \mathbf{X}) &= \pi_1^T \mathbf{X}, \\ y^i(\pi_3^T \mathbf{X}) &= \pi_2^T \mathbf{X} \end{aligned}$$

Calibration with a Rig

- Recover projection matrix $\Pi = [KR, KT] = [R', T']$

$$\Pi^s = [\pi_{11}, \pi_{21}, \pi_{31}, \pi_{12}, \pi_{22}, \pi_{32}, \pi_{13}, \pi_{23}, \pi_{33}, \pi_{14}, \pi_{24}, \pi_{34}]^T$$

$$\min \|M\Pi^s\|^2 \quad \text{subject to} \quad \|\Pi^s\|^2 = 1$$

Again singular value decomposition

- Factor the KR into $R \in SO(3)$ and K using QR decomposition
- Solve for translation $T = K^{-1}T'$

Binocular Stereo

Binocular Stereo

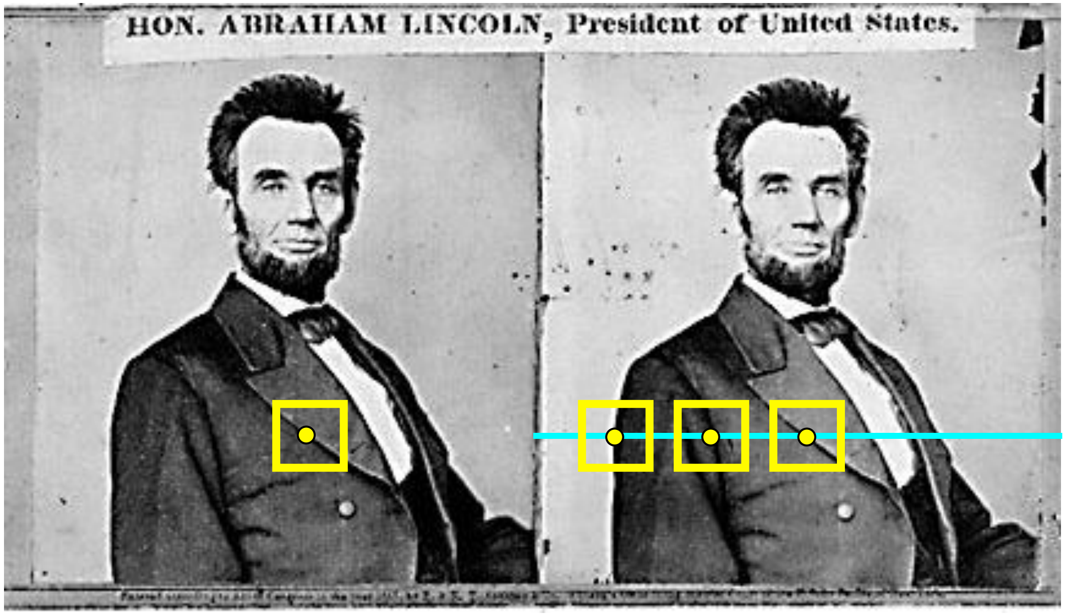
- Given a calibrated binocular stereo pair, fuse it to produce a depth image

image 1

image 2

Dense depth map

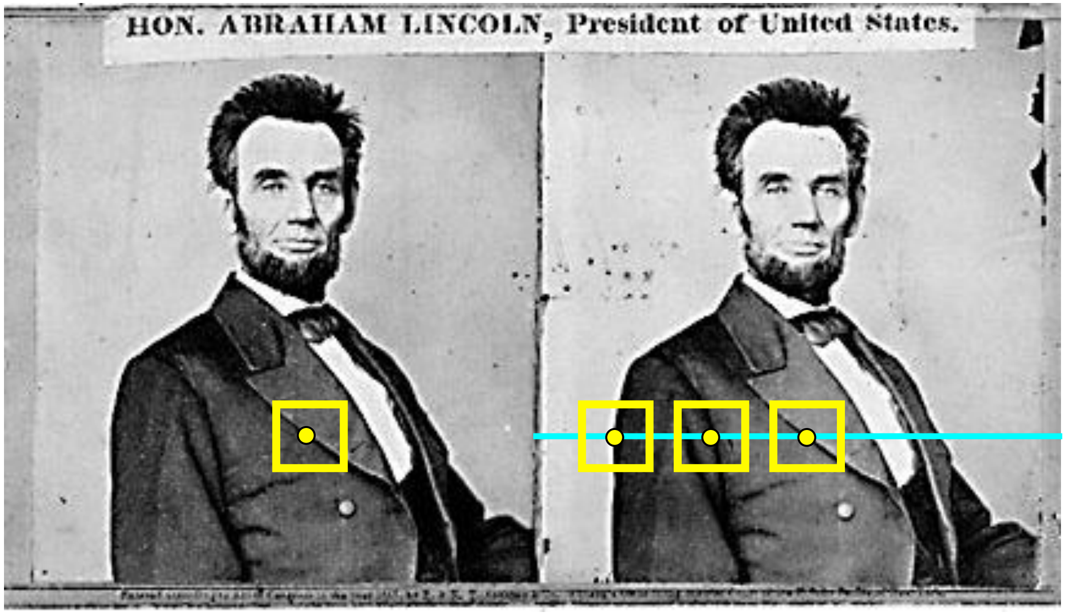
Basic Stereo Matching Algorithm



- For each pixel in the first image
  - Find corresponding epipolar line in the right image
  - Examine all pixels on the epipolar line and pick the best match
  - Triangulate the matches to get depth information

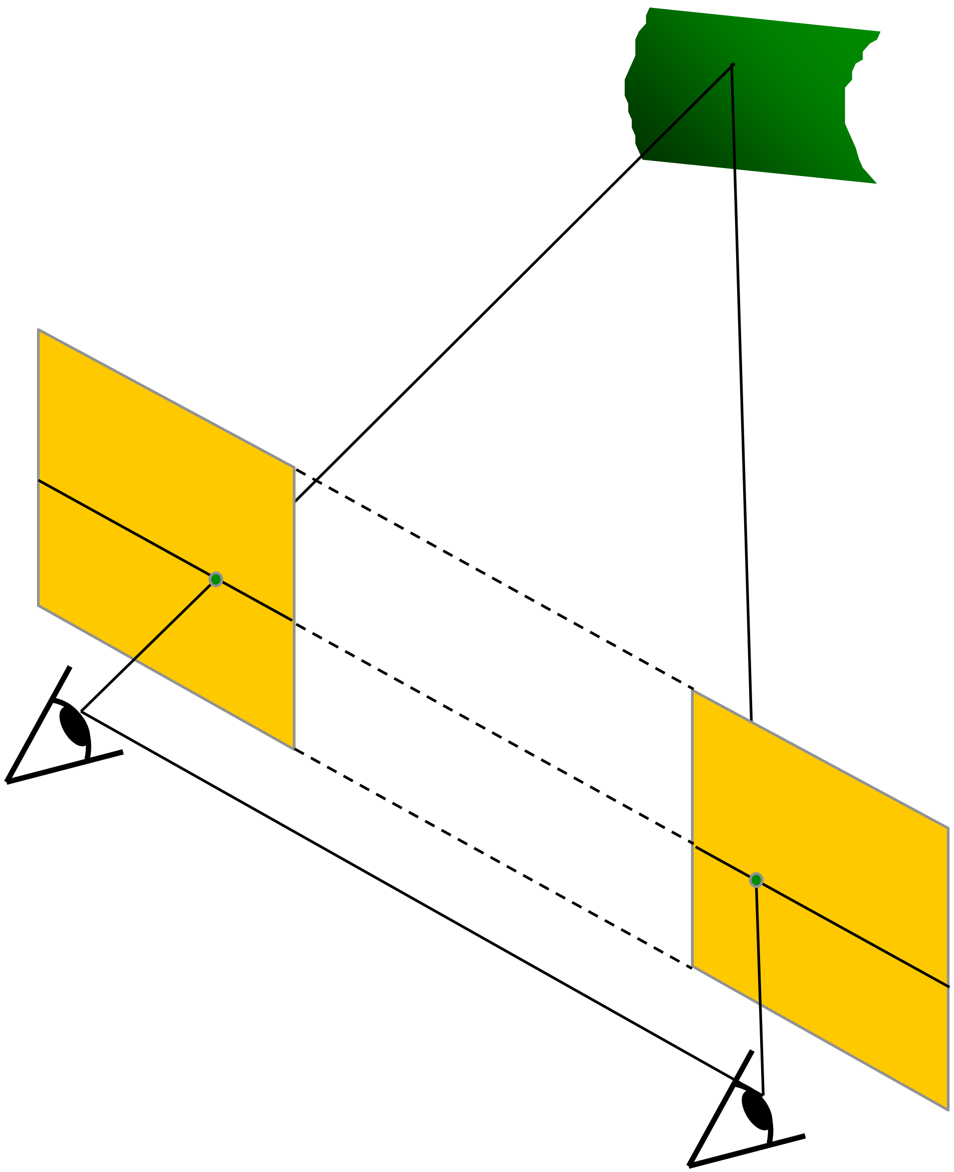
- Simplest case: epipolar lines are corresponding scanlines
  - When does this happen?

Basic stereo matching algorithm



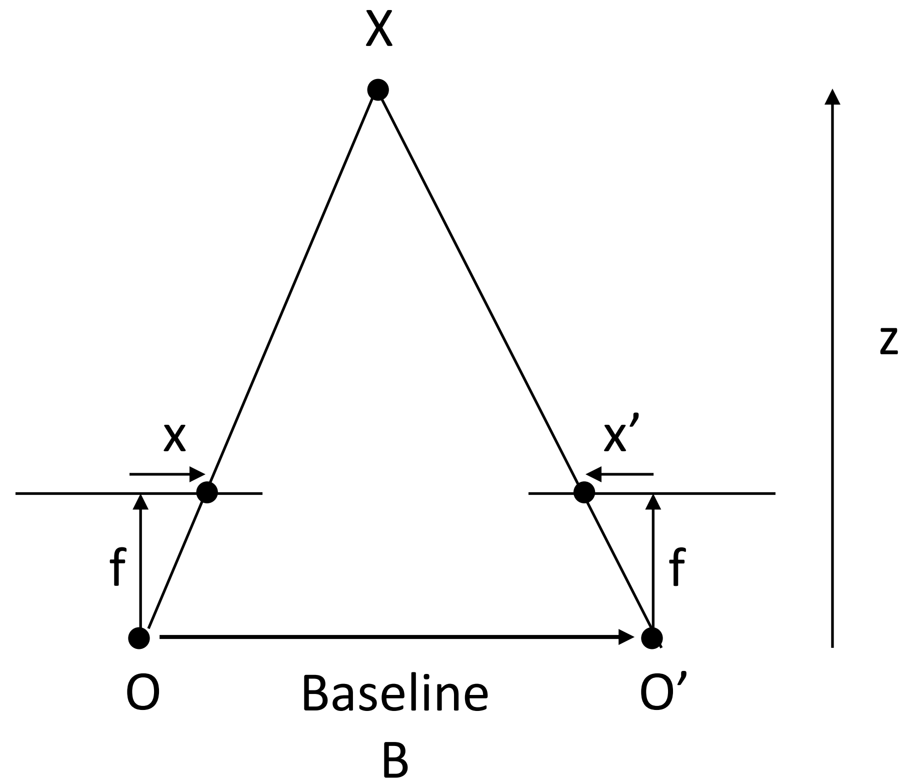
- For each pixel in the first image
  - Find corresponding epipolar line in the right image
  - Examine all pixels on the epipolar line and pick the best match
  - Triangulate the matches to get depth information
- Simplest case: epipolar lines are corresponding scanlines
  - When does this happen?

Simplest Case: Parallel Images



- Image planes of cameras are parallel to each other and to the baseline
- Camera centers are at same height
- Focal lengths are the same
- Then, epipolar lines fall along the horizontal scan lines of the images

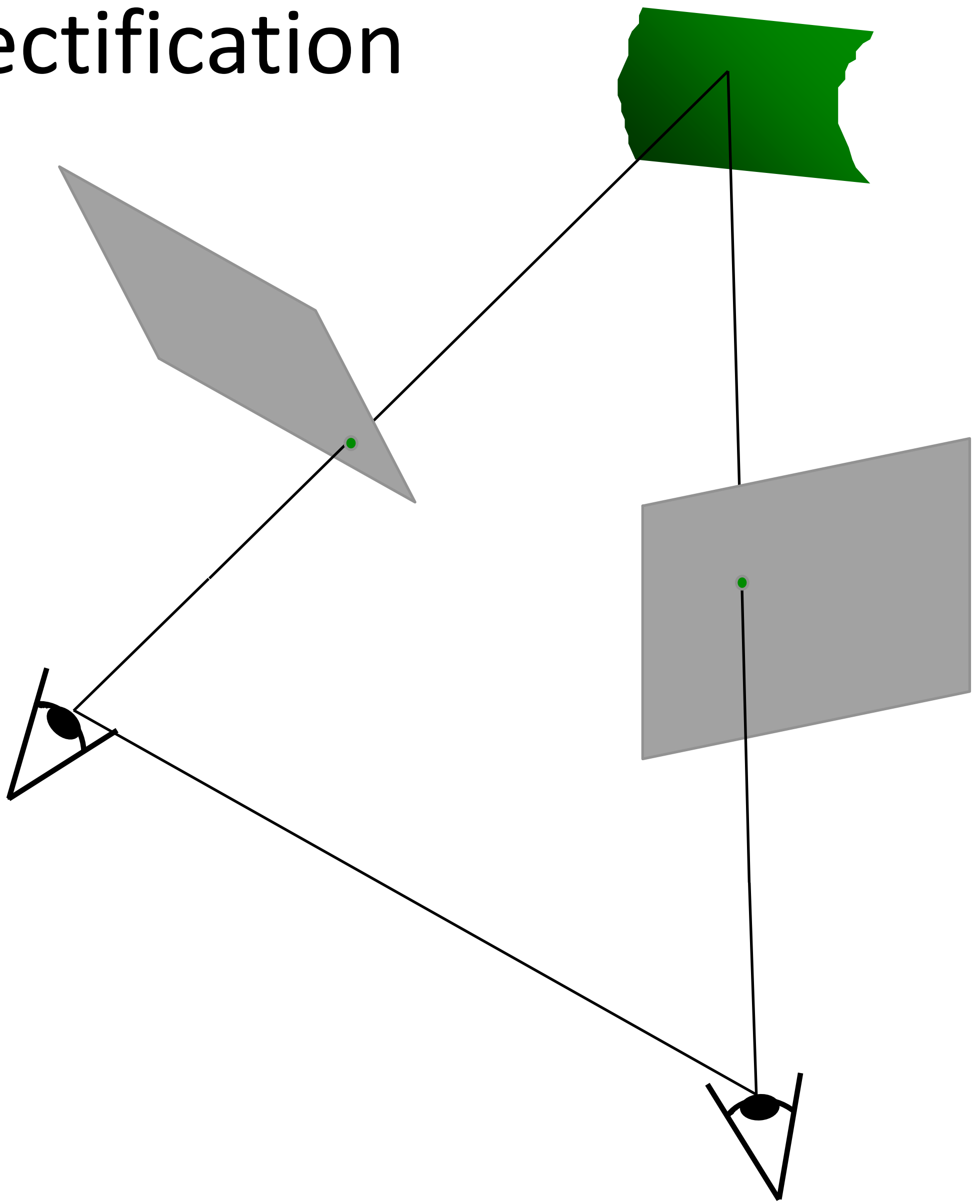
Depth from Disparity



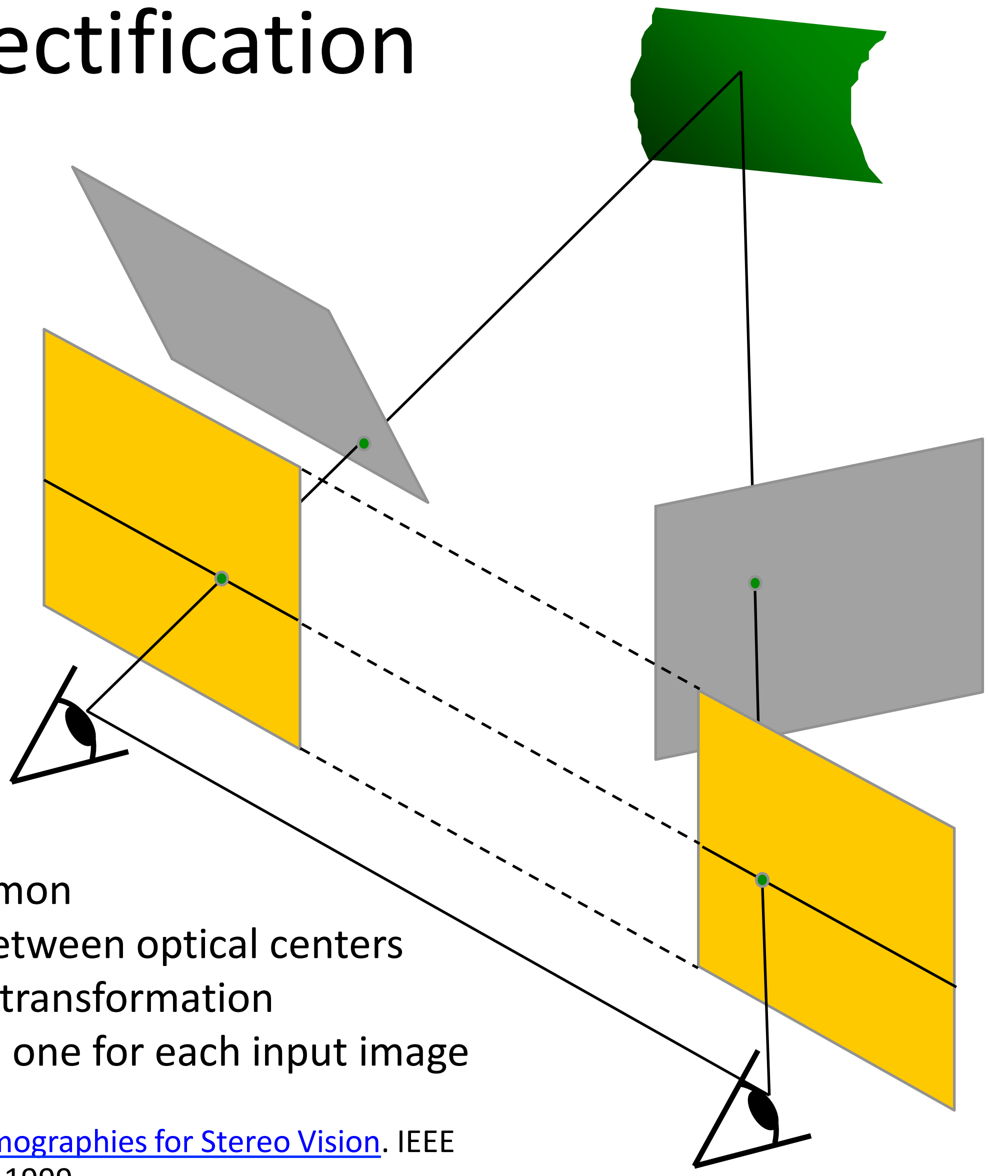
$$disparity = x - x' = \frac{B \times f}{z}$$

Disparity is inversely proportional to depth!

Stereo Image Rectification



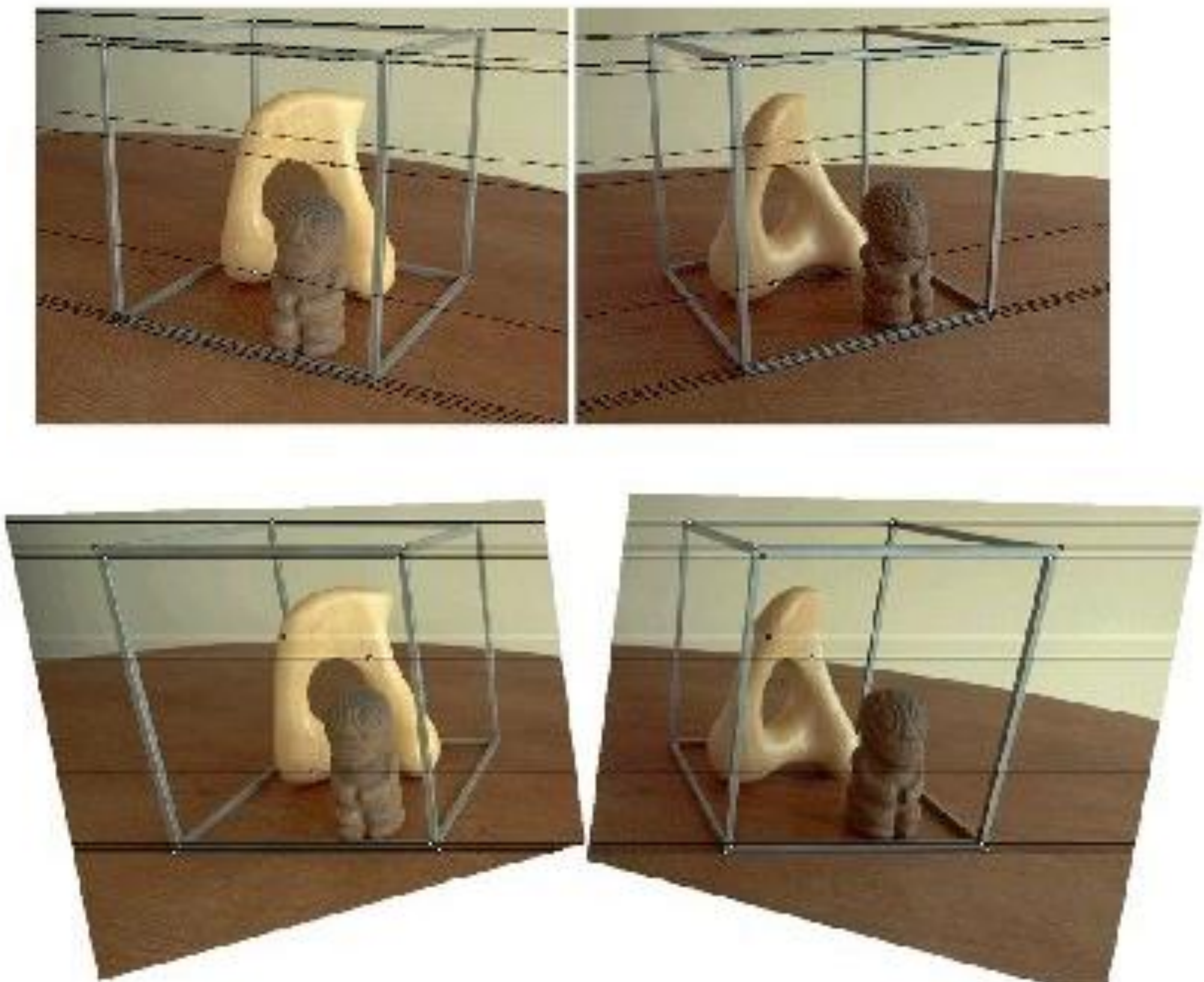
Stereo Image Rectification



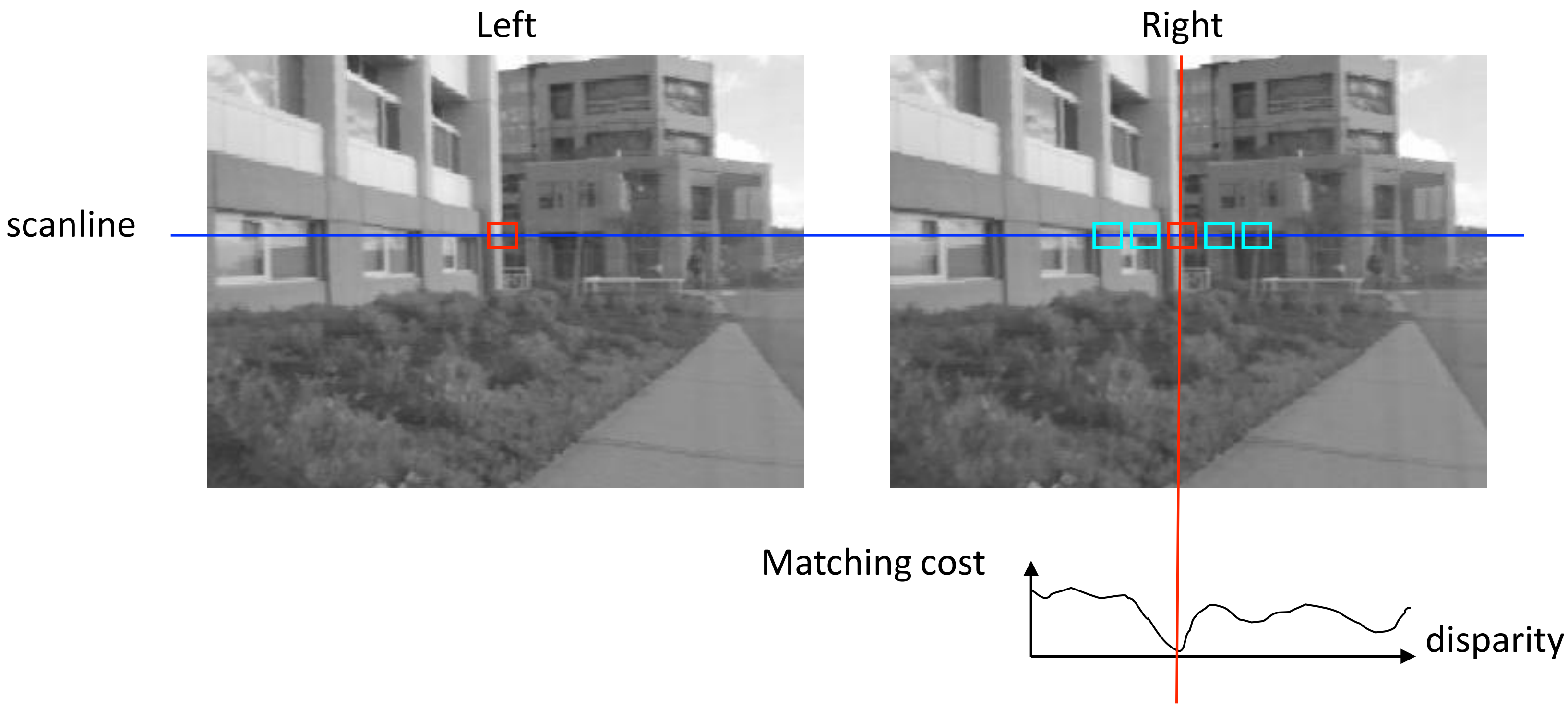
- reproject image planes onto a common
- plane parallel to the line between optical centers
- pixel motion is horizontal after this transformation
- two homographies (3x3 transform), one for each input image reprojection

➢C. Loop and Z. Zhang. [Computing Rectifying Homographies for Stereo Vision](). IEEE Conf. Computer Vision and Pattern Recognition, 1999.

Rectification Example

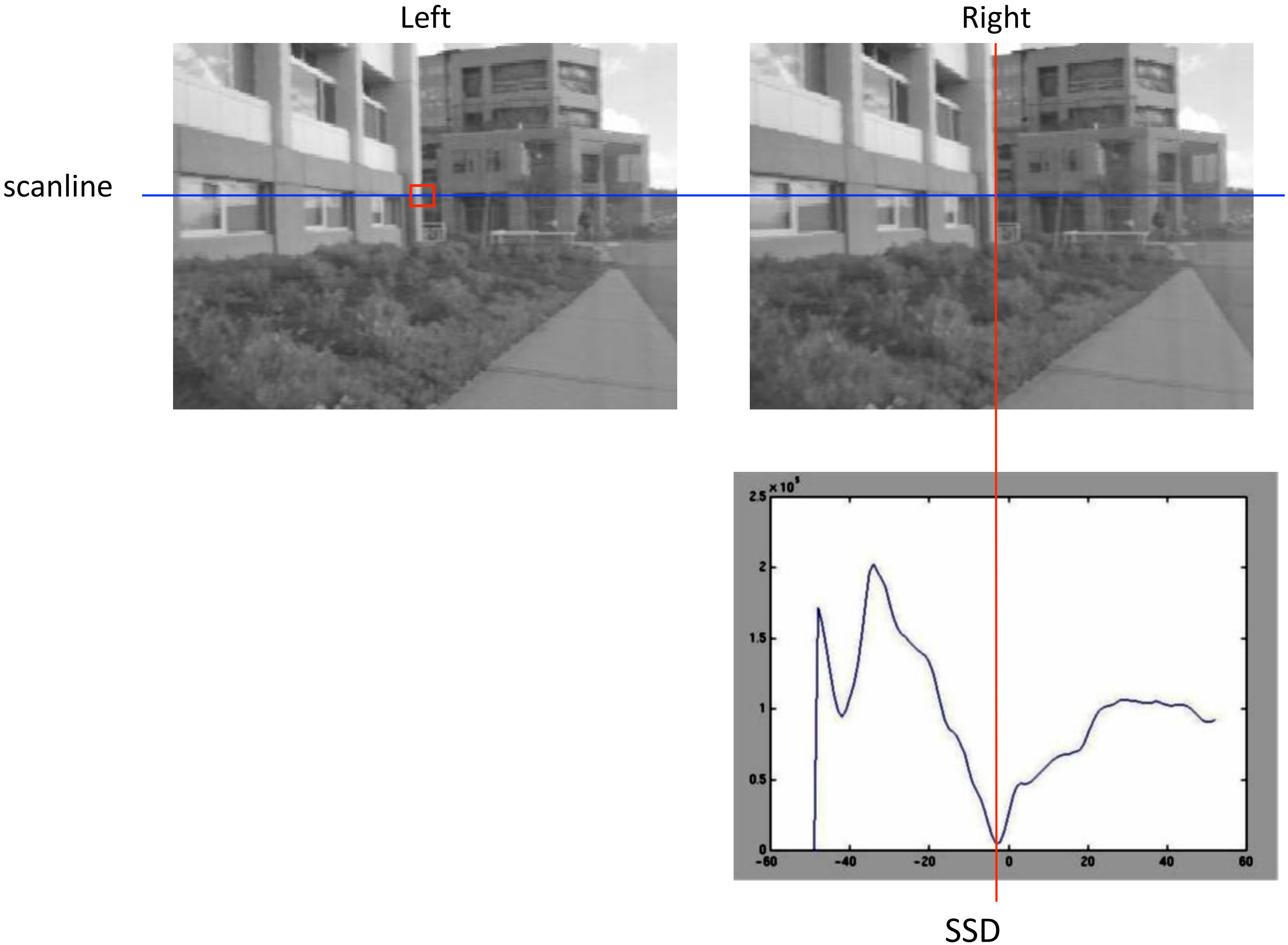


Correspondence search



- Slide a window along the right scanline and compare contents of that window with the reference window in the left image
- Matching cost: SSD or normalized correlation

Correspondence search

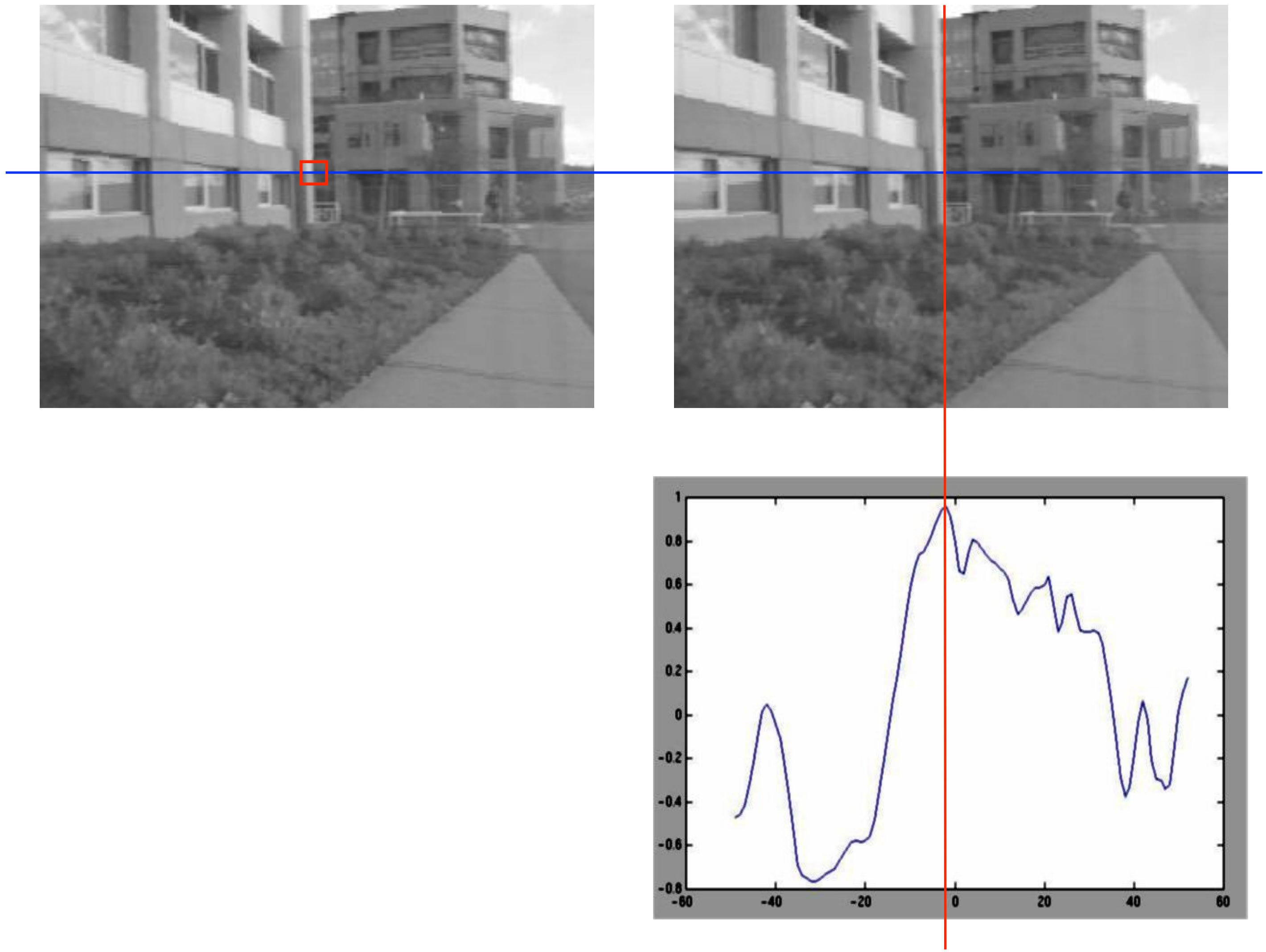


Correspondence search

Left

Right

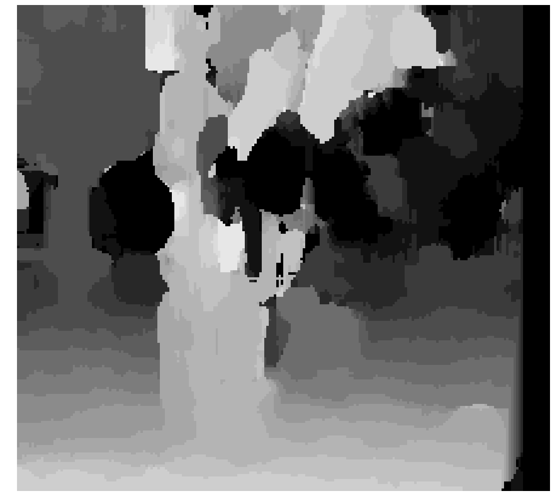
scanline



Norm. corr

Effect of window size

W = 3



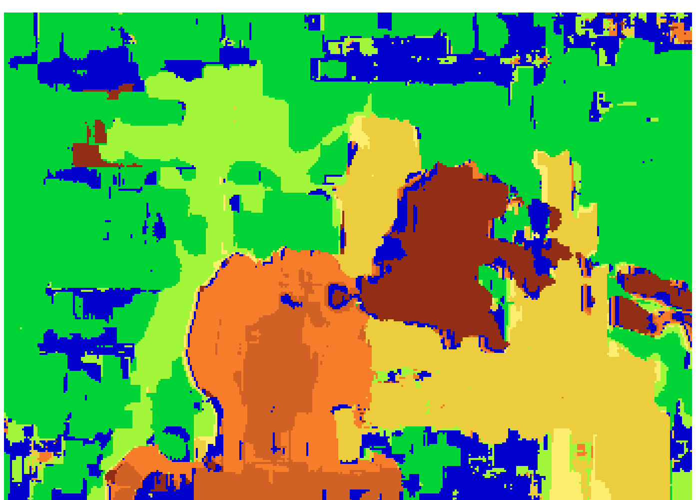
W = 20

- Smaller window
  - \+ More detail
  - – More noise
- Larger window
  - \+ Smoother disparity maps
  - – Less detail

Results with window search

Data

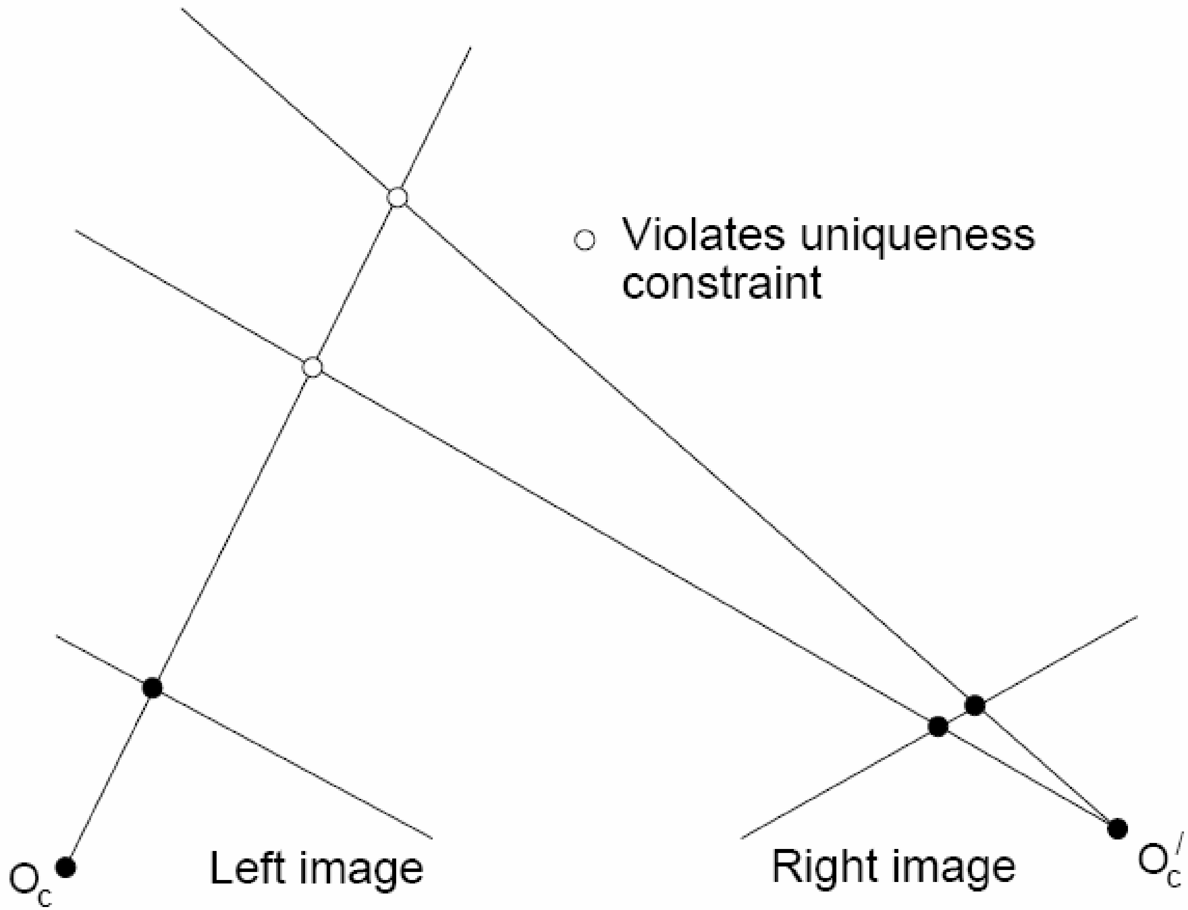
Window-based matching



Ground truth

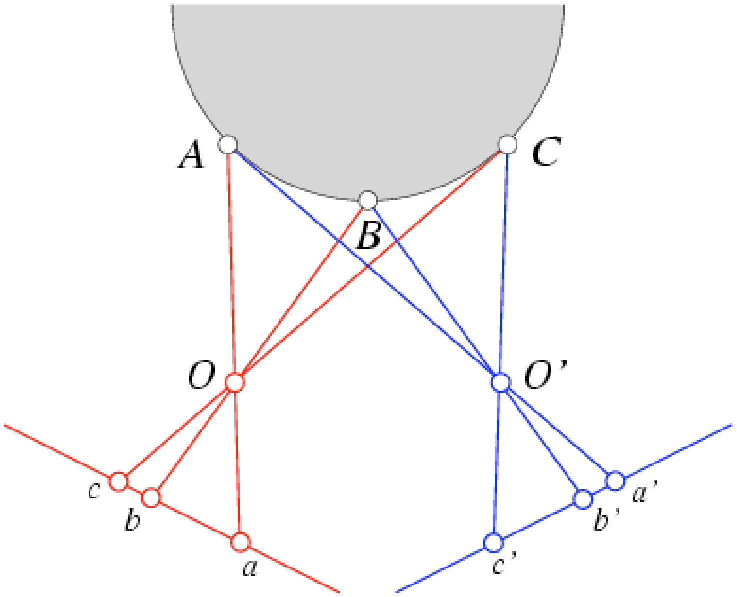
Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image



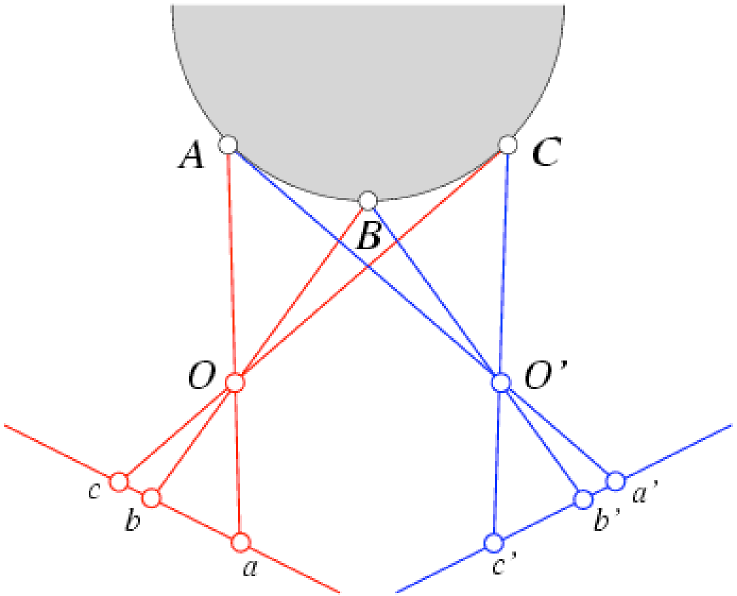
Non-local constraints

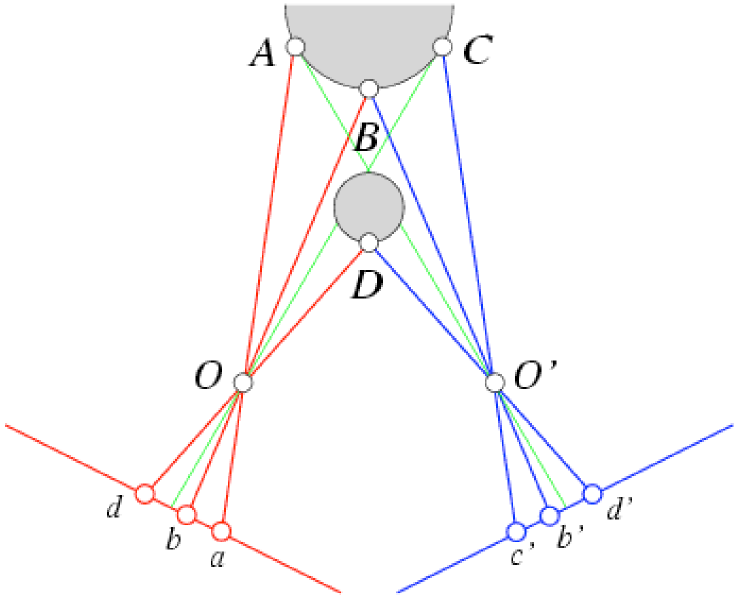
- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views



Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views





Ordering constraint doesn't hold

Consistency Constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views
- Smoothness
  - We expect disparity values to change slowly (for the most part)

MRF Formulation:

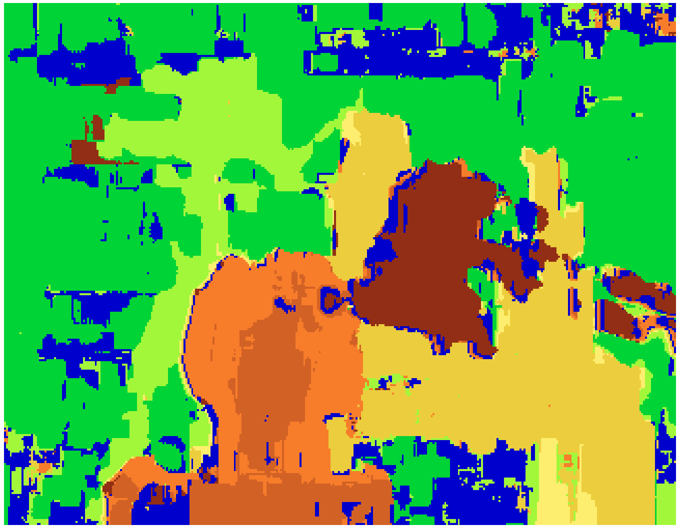
$$E(d) = E_d(d) + \lambda E_s(d)$$

Pixel matching score

Consistency Scores

Comparsion

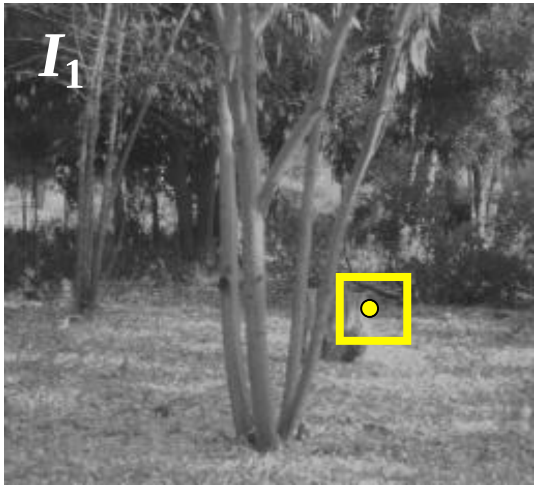
Window-Based Search:

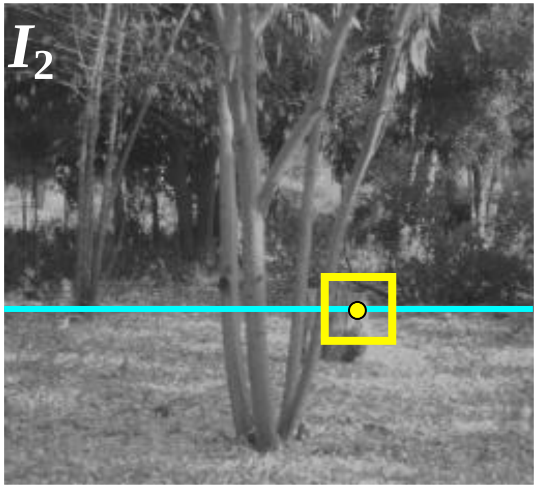


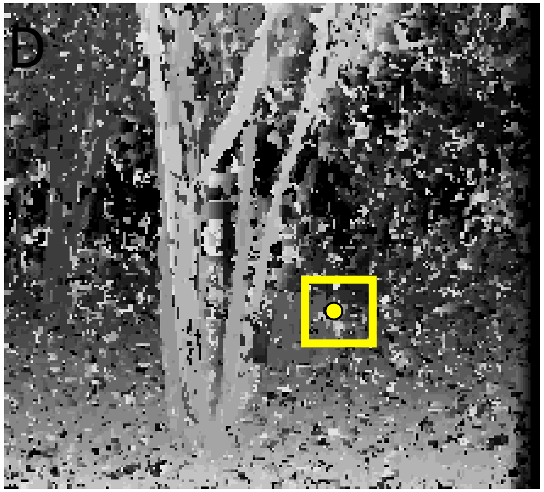
Graph Cut:

Ground Truth

Stereo matching as energy minimization





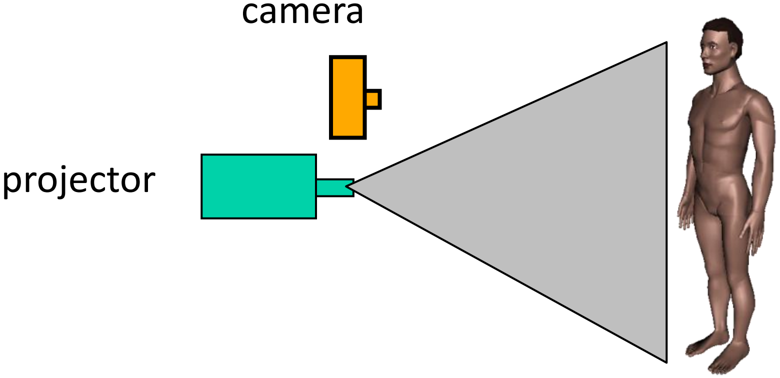


- Graph-cuts can be used to minimize such energy

  Y. Boykov, O. Veksler, and R. Zabih, [Fast Approximate Energy Minimization via Graph Cuts](), PAMI 2001

Active stereo with structured light

- Project “structured” light patterns onto the object
  - Simplifies the correspondence problem
  - Allows us to use only one camera



L. Zhang, B. Curless, and S. M. Seitz. [Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming.]() *3DPVT* 2002

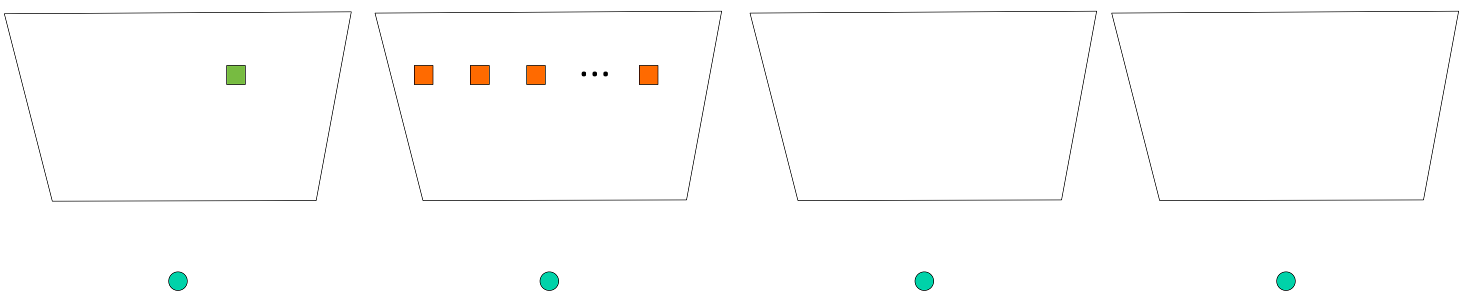
Kinect: Structured infrared light

http://bbzippo.wordpress.com/2010/11/28/kinect-in-infrared/

Multi-Baseline Stereo

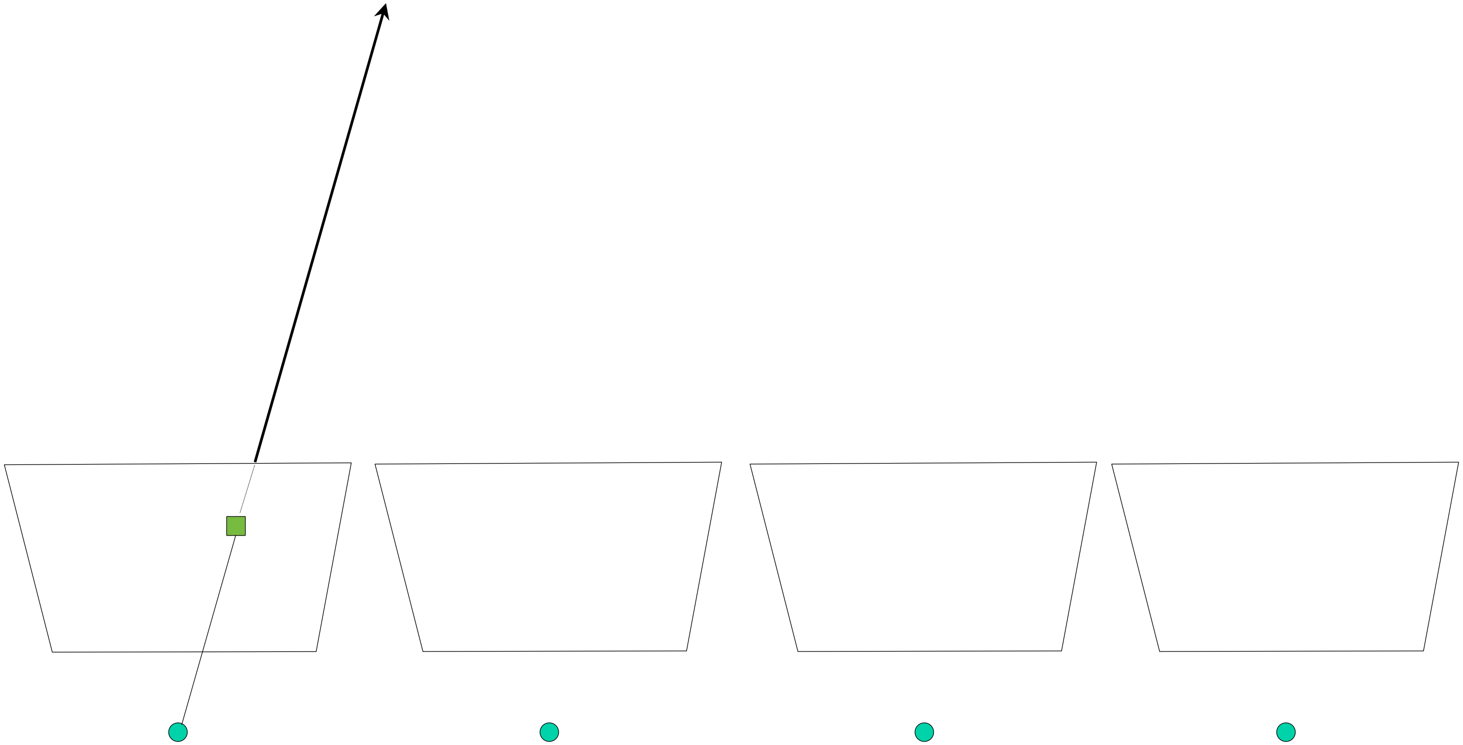
Same formulation with more images

- Change label from disparity to depth
- Change $E_d(d)$ by using more images



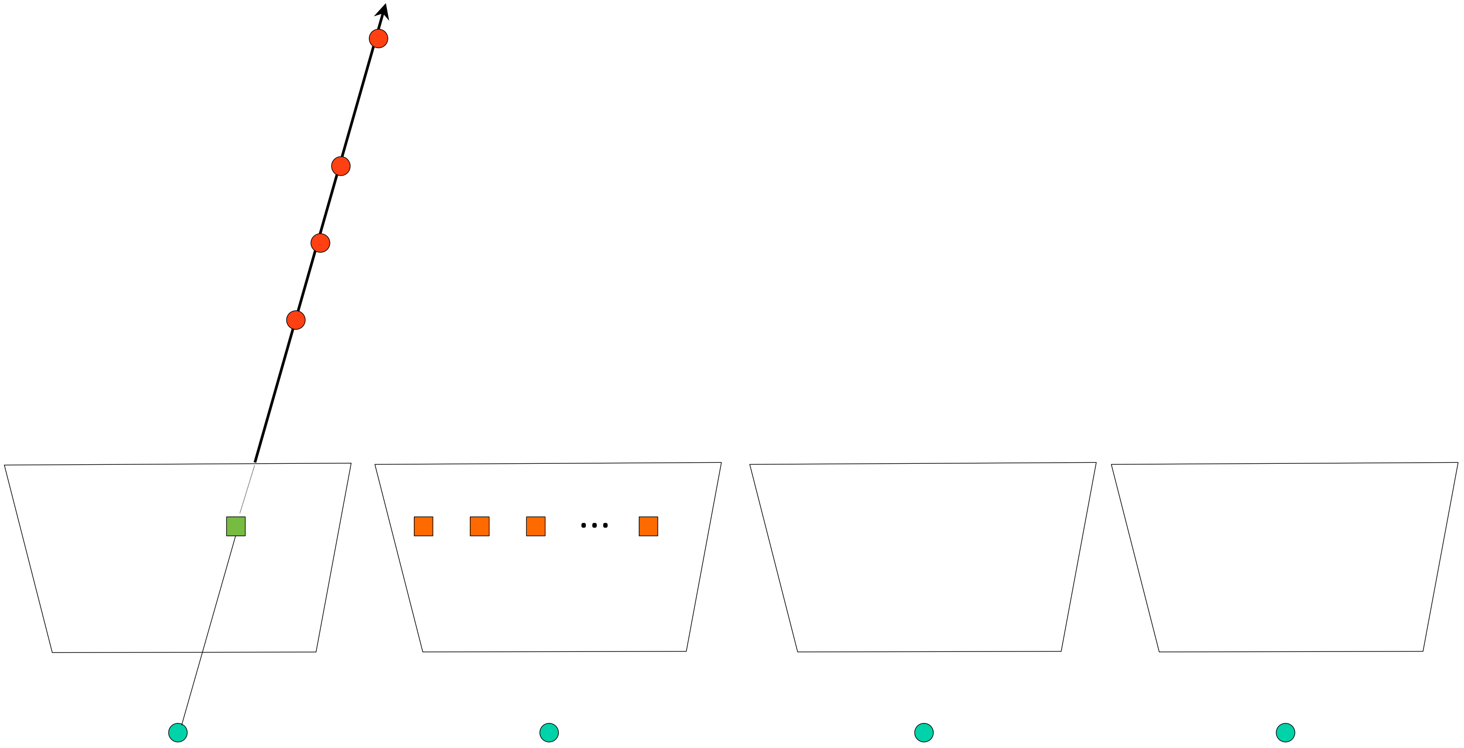
Same formulation with more images

- Change label from disparity to depth
- Change $E_d(d)$ by using more images



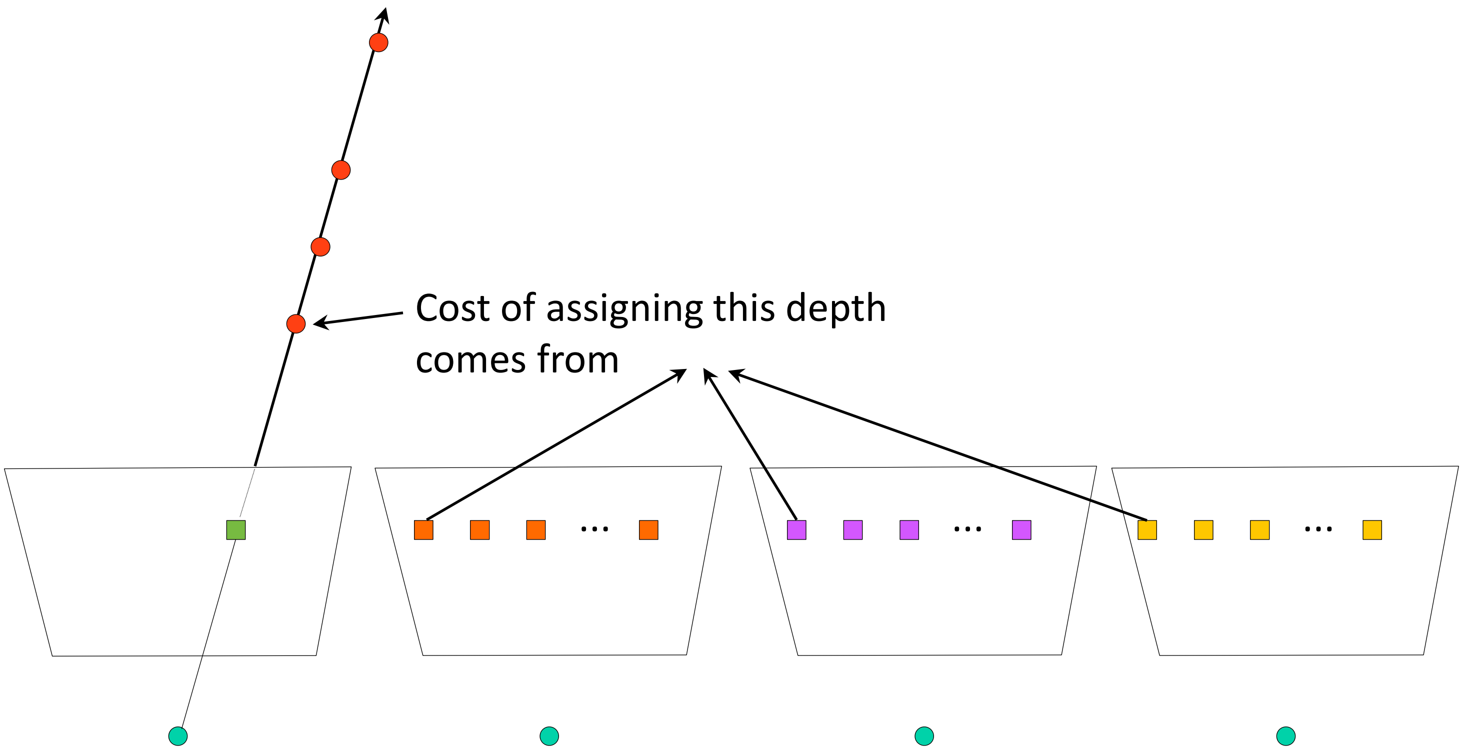
Same formulation with more images

- Change label from disparity to depth
- Change $E_d(d)$ by using more images

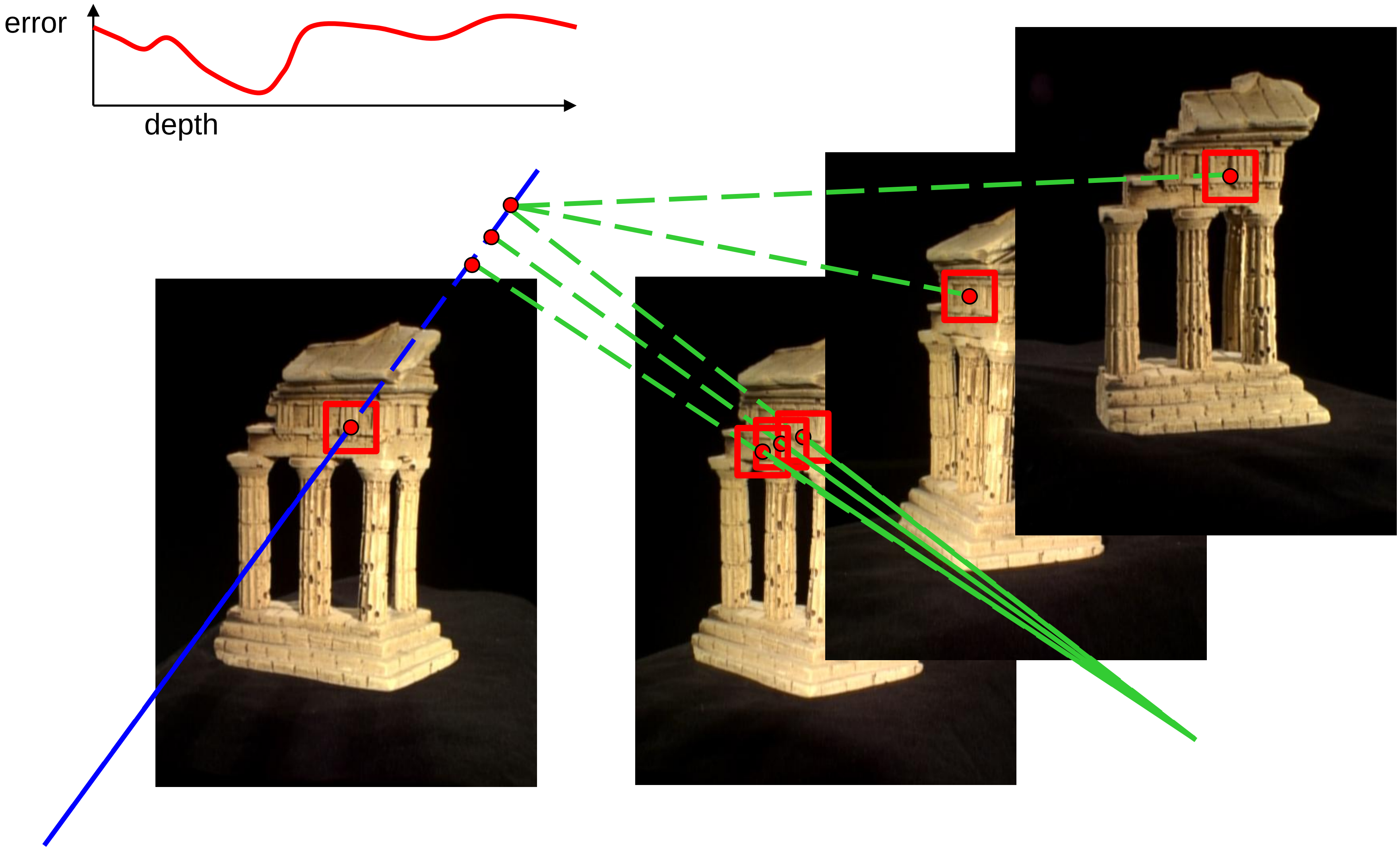


Same formulation with more images

- Change label from disparity to depth
- Change $E_d(d)$ by using more images



Stereo: Basic Idea



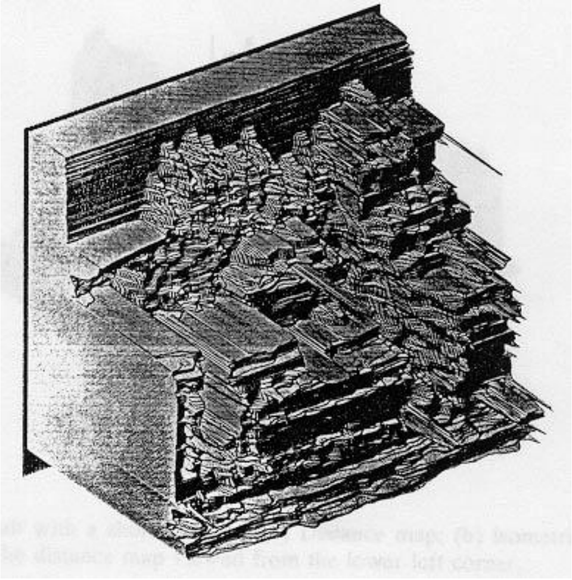
Multiple-Baseline Stereo Results

[Okutomi and Kanade' 93]

I1

I2

I10



Mesh Reconstruction

Merging Depth Maps

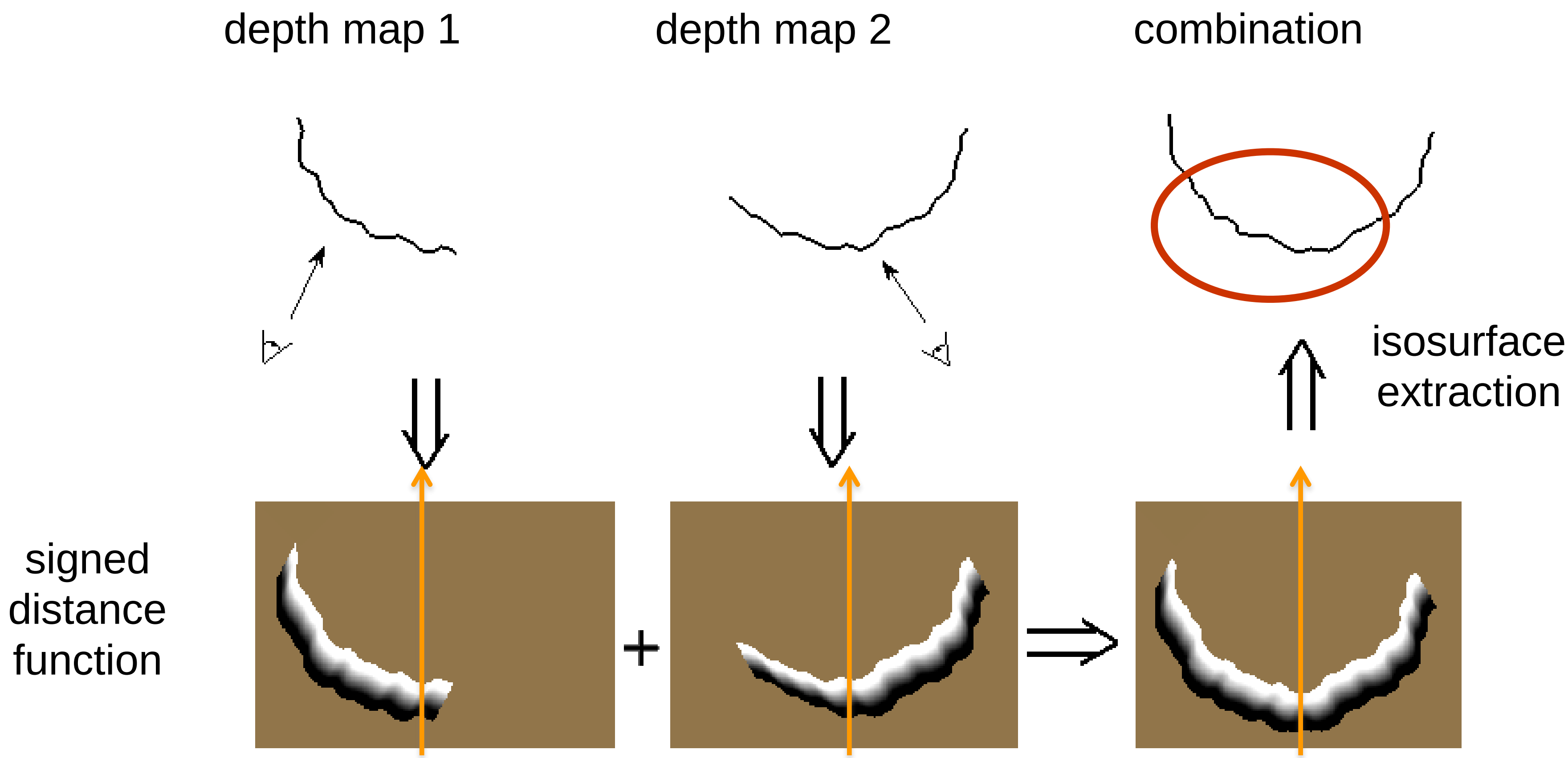
vrip [Curless and Levoy 1996]

- compute weighted average of depth maps

set of depth maps
(one per view)

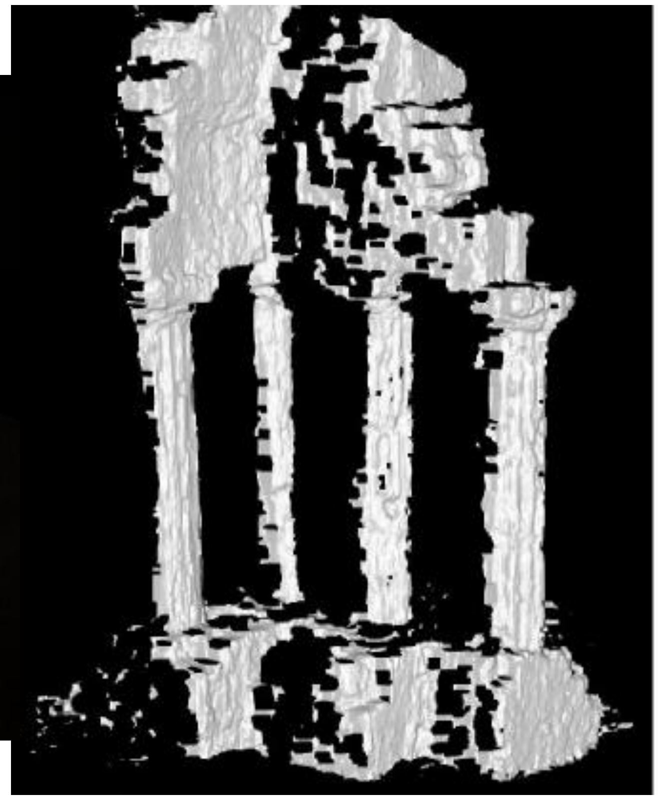
merged surface
mesh

VRIP

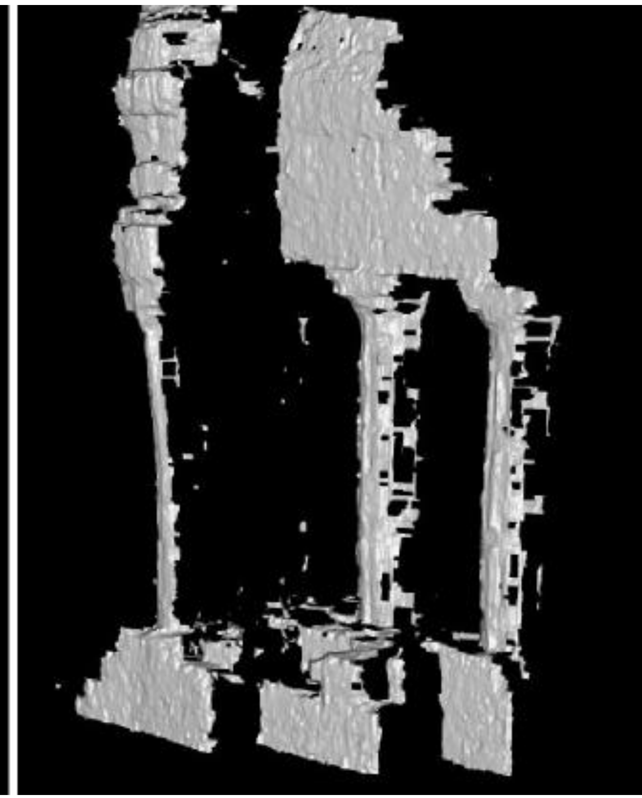

depth map 1
depth map 2
combination
isosurface
extraction
signed
distance
function
+

Depthmap Merging

Depthmap 1



Depthmap 2



Merging Depth Maps: Temple Model

[Goesele et al. 06]

input image

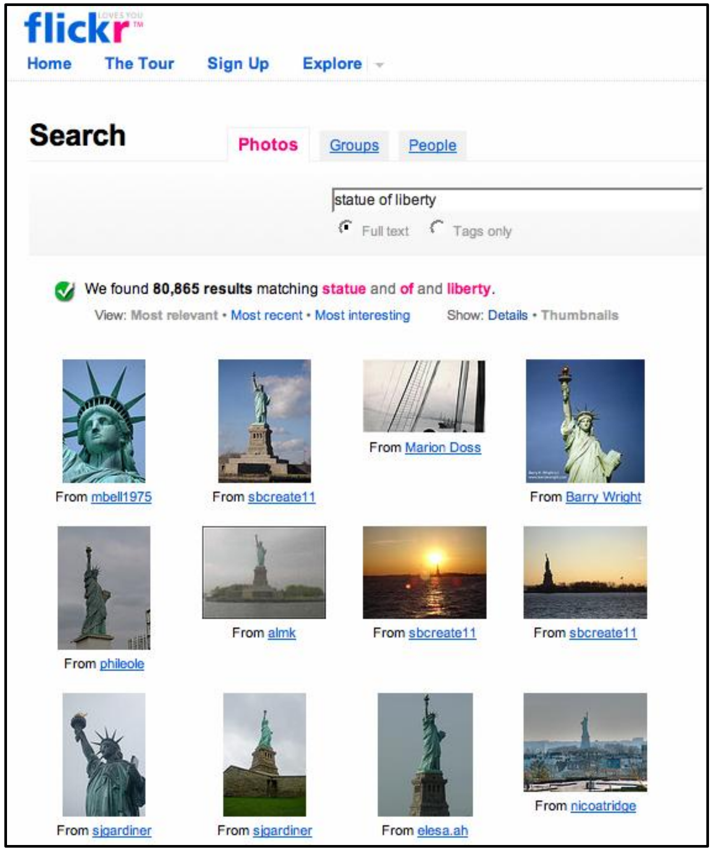
317 images
(hemisphere)

ground truth model

State-of-The-Art

Multi-View Stereo from Internet Collections

[Goesele et al. 07]



Challenges

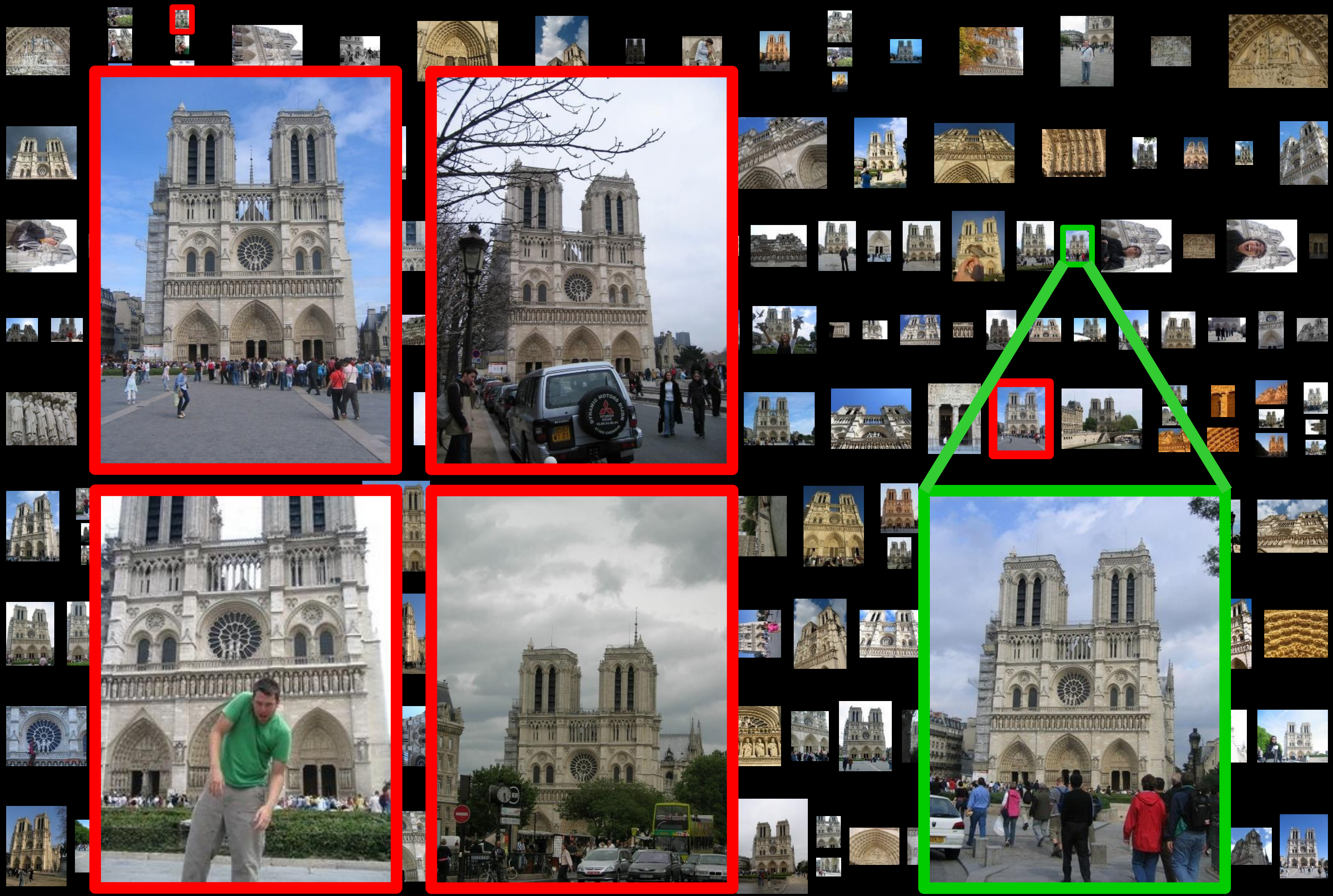
- Appearance variation

- Resolution

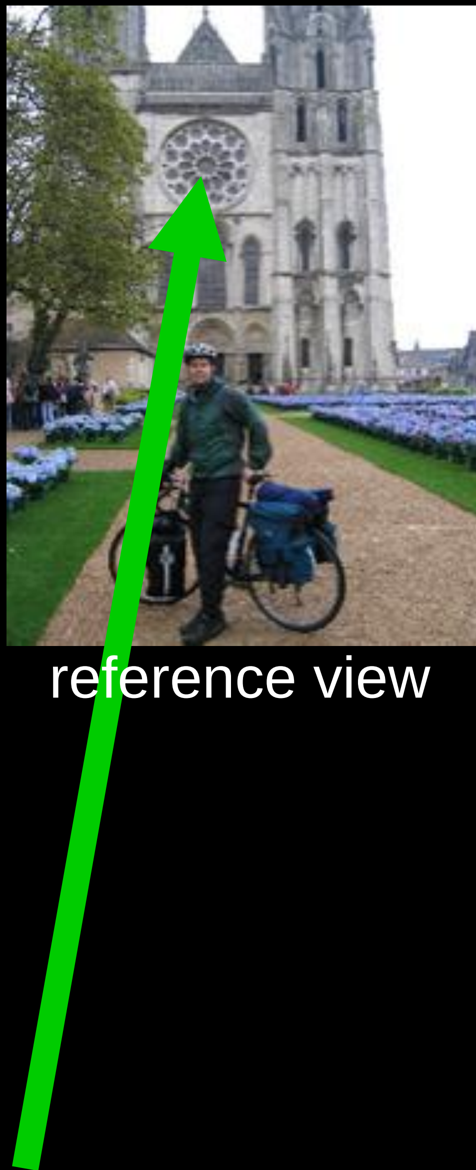
- Massive collections

82754 results for photos matching **notre** and **dame** and **paris**

Law of Nearest Neighbors



206 *Flickr* images taken by 92 photographers

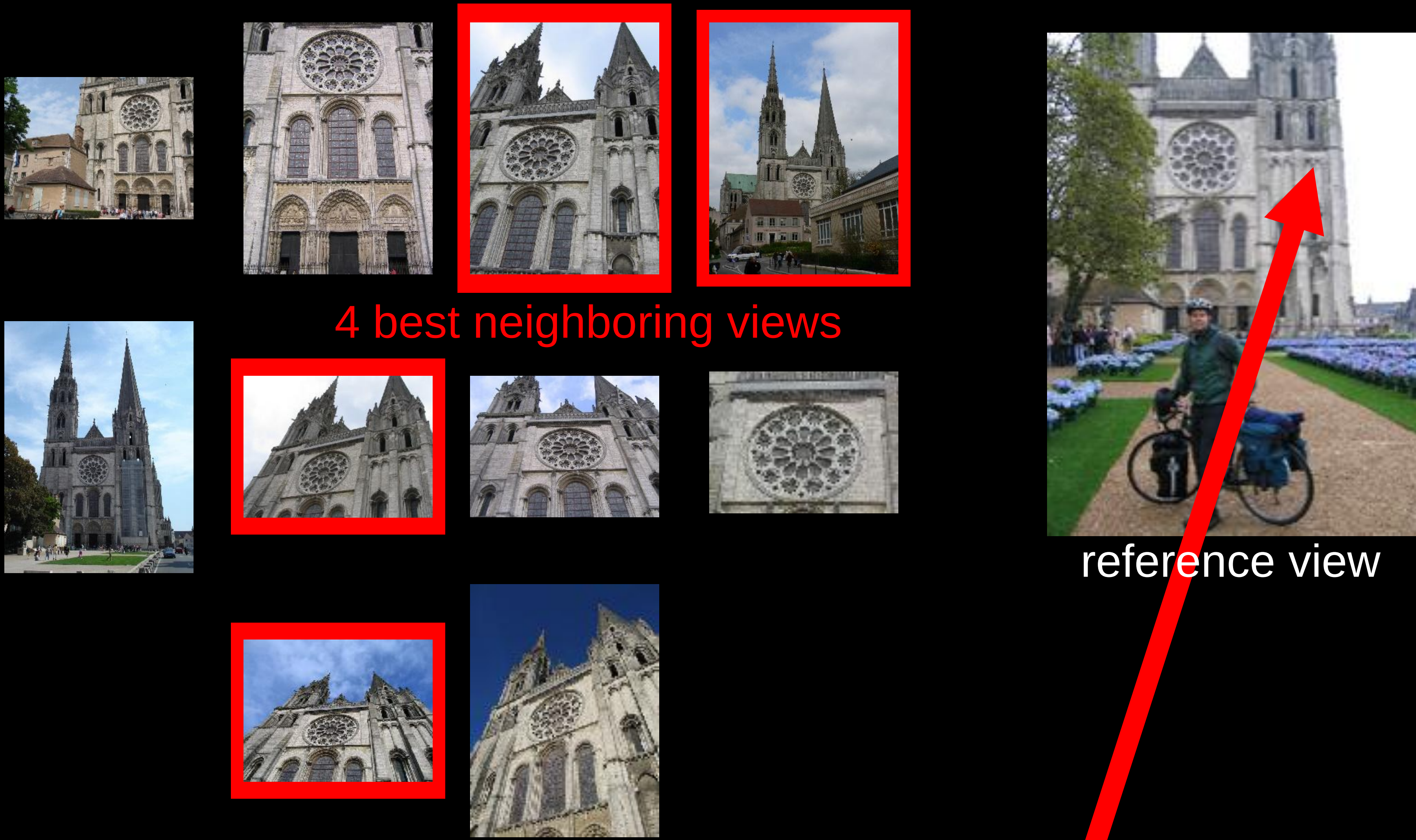


Local view selection

- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines

Local view selection

- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines

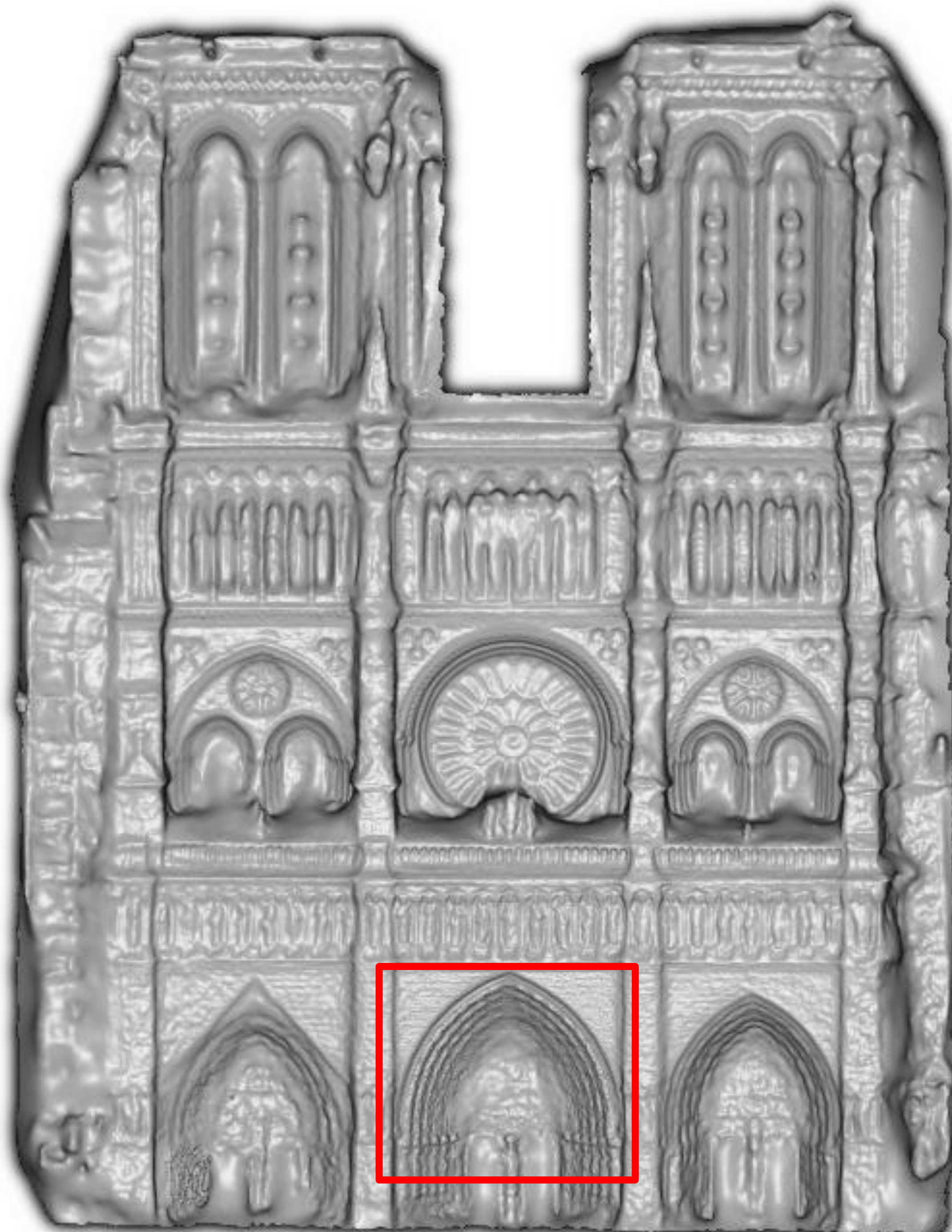


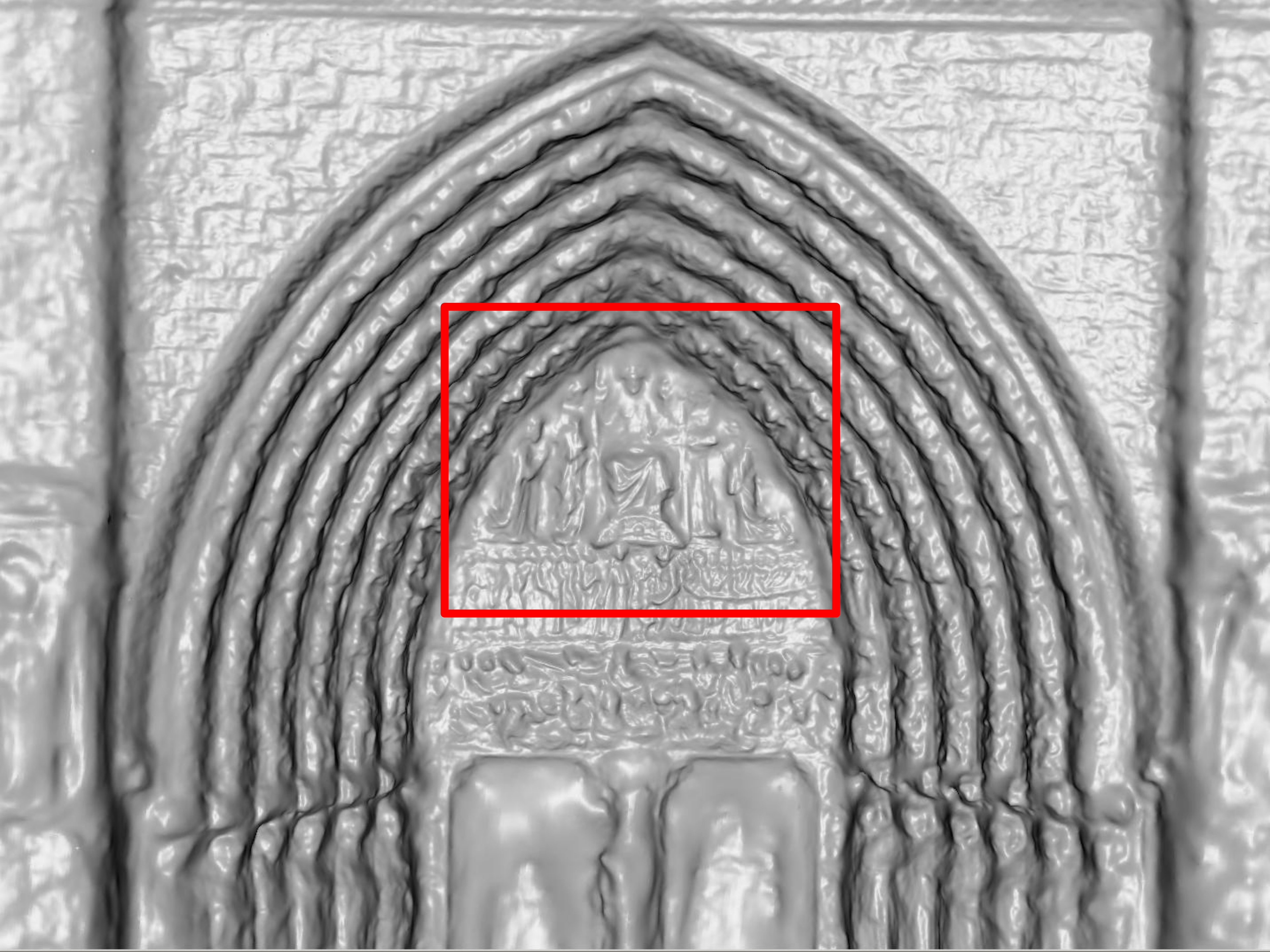
Local view selection

- Automatically select neighboring views for each point in the image
- Desiderata: good matches AND good baselines

Notre Dame de Paris

653 images
313 photographers

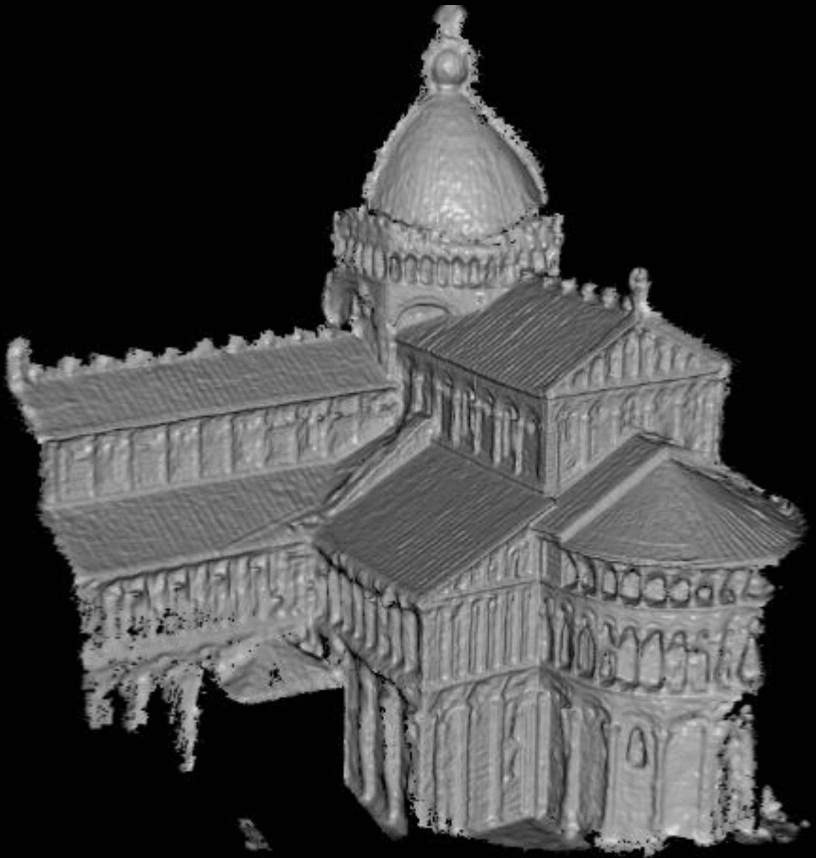




129 *Flickr* images taken by 98 photographers

merged model of Venus de Milo

56 *Flickr* images taken by 8 photographers



merged model of Pisa Cathedral

Accuracy compared to laser scanned model:
90% of points within *0.25%* of ground truth

How can Deep Learning Help?