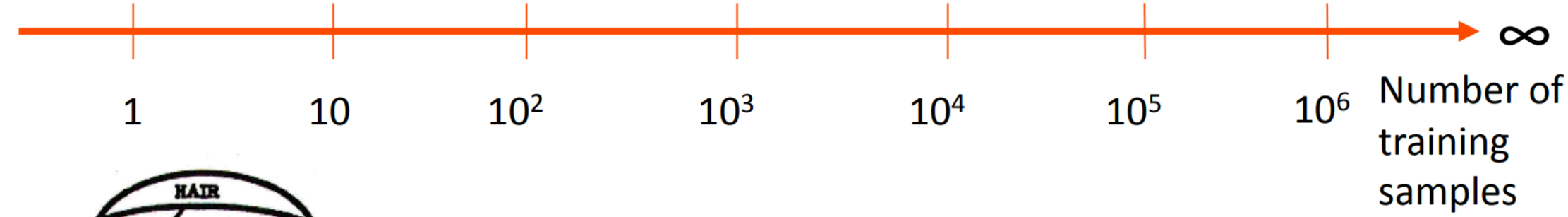
Finishing Up Object Recognition

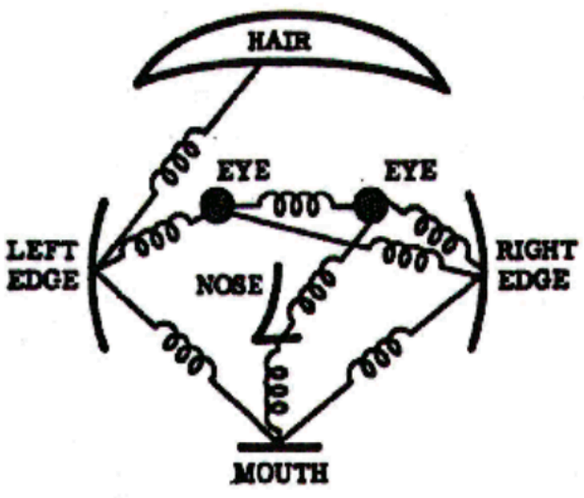
Computer Vision
Fall 2019
Columbia University

Two Extremes of Vision

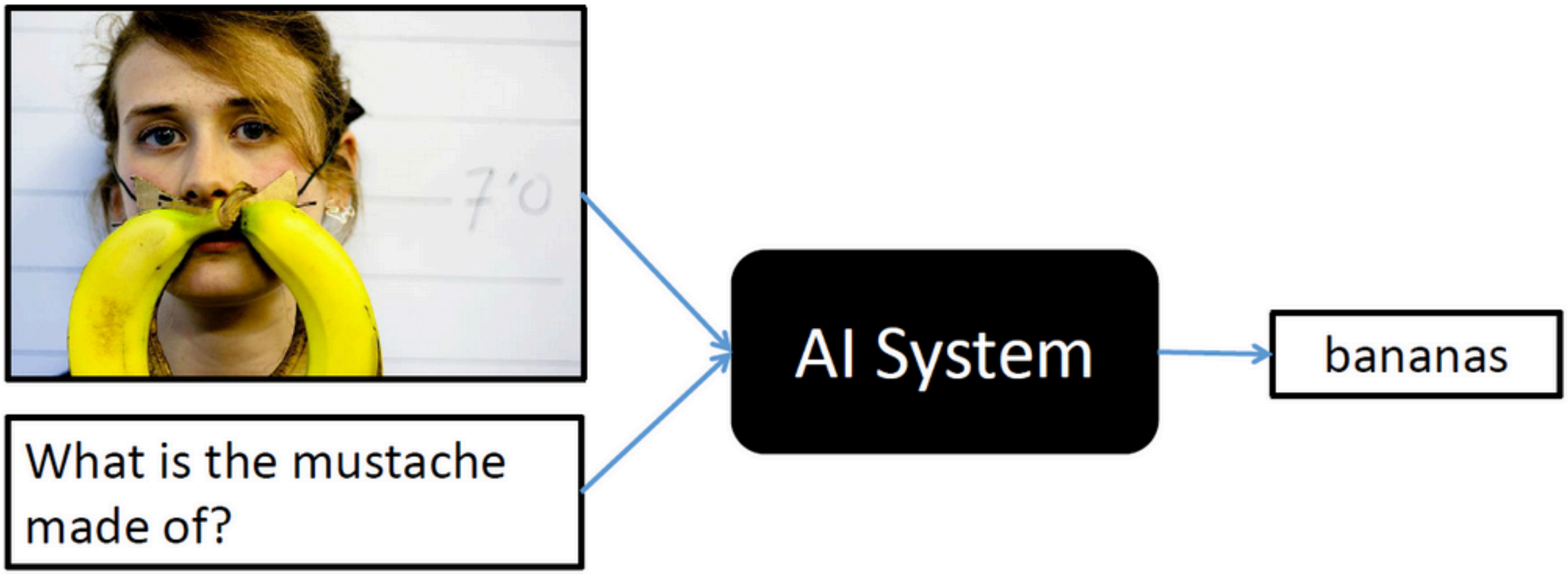
Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences





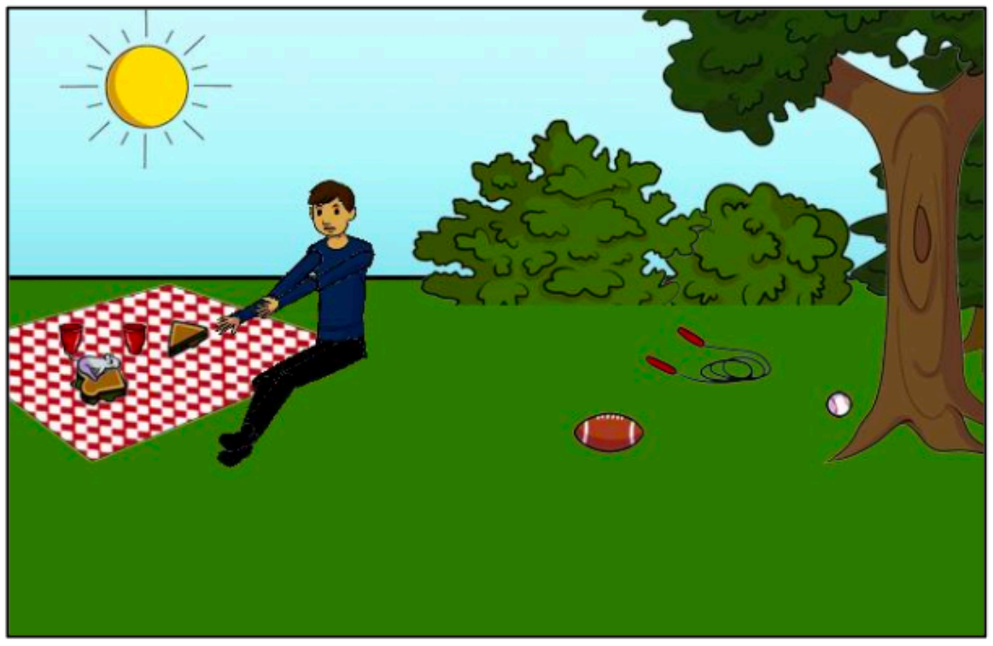
Slide credit: Aude Oliva



http://www.visualqa.org/challenge.html

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

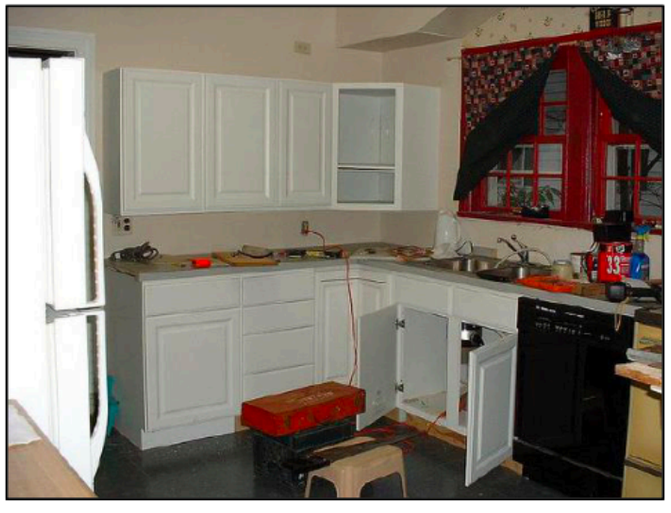


Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

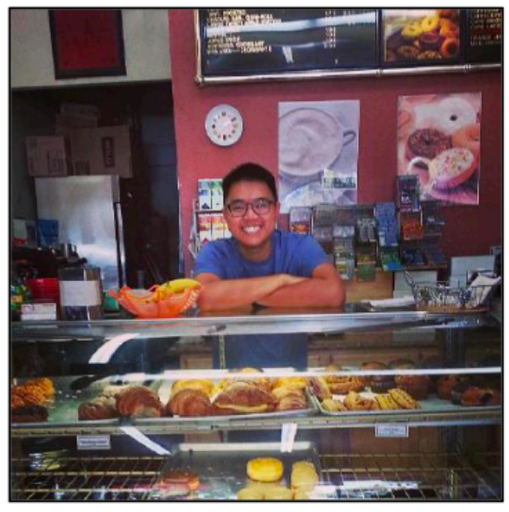
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Questions and answers collected with Amazon Mechanical Turk

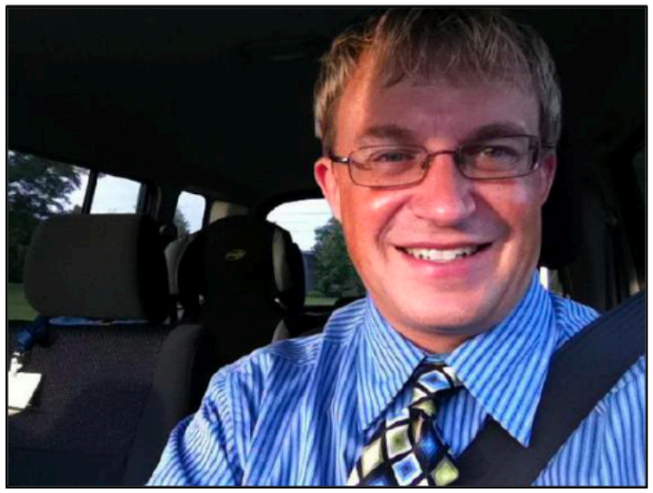


Is something under the sink broken?	yes yes yes	no no no
What number do you see?	33 33 33	5 6 7

Can you park here?	no no no	no no yes
What color is the hydrant?	white and orange white and orange white and orange	red red yellow



What kind of store is this?	bakery bakery pastry	art supplies grocery grocery
Is the display case as full as it could be?	no no no	no yes yes



Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes

Has the pizza been baked?	yes yes yes	yes yes yes
What kind of cheese is topped on this pizza?	feta feta ricotta	mozzarella mozzarella mozzarella

How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round

Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

Architecture

Words

- Need ways to compare words

Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'	'man'
'bench'	'woman'
'chair'	'child'
'deck chair'	'teenager'
'ottoman'	'girl'
'seat'	'boy'
'stool'	'baby'
'swivel chair'	'daughter'
'loveseat'	'son'
…	…

word2vec

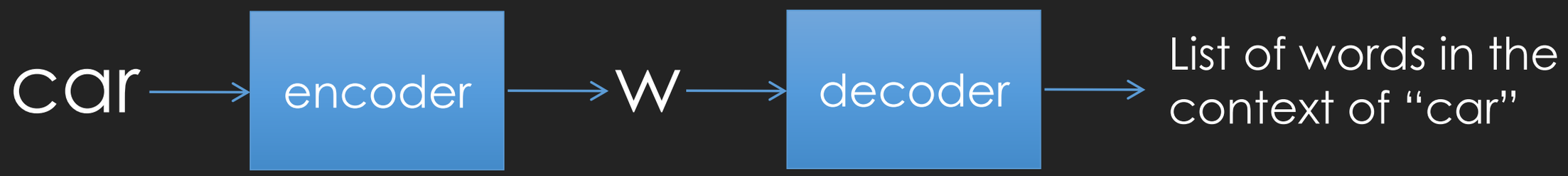
I parked the car in a nearby street. It is a red car with two doors, …

I parked the vehicle in a nearby street…

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

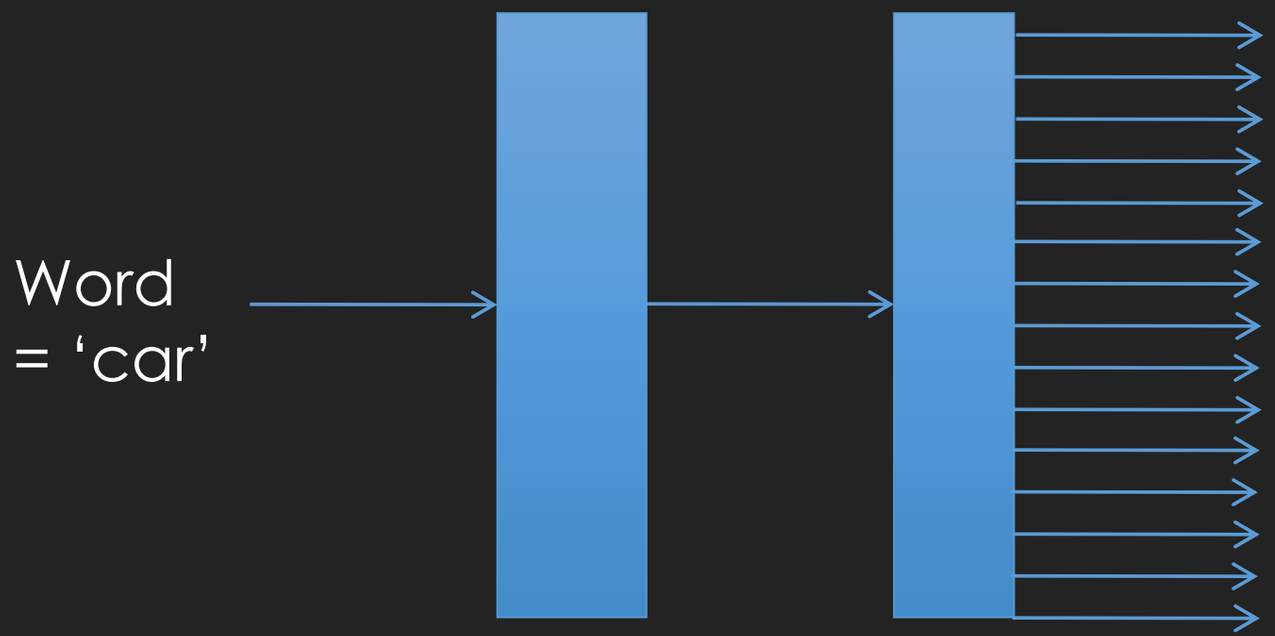
word2vec

I parked the car in a nearby street. It is a red car with two doors, …



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec



Output prob. That each word is in the context of the input word

Hidden layer Soft-max classifier

Encoder Decoder

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

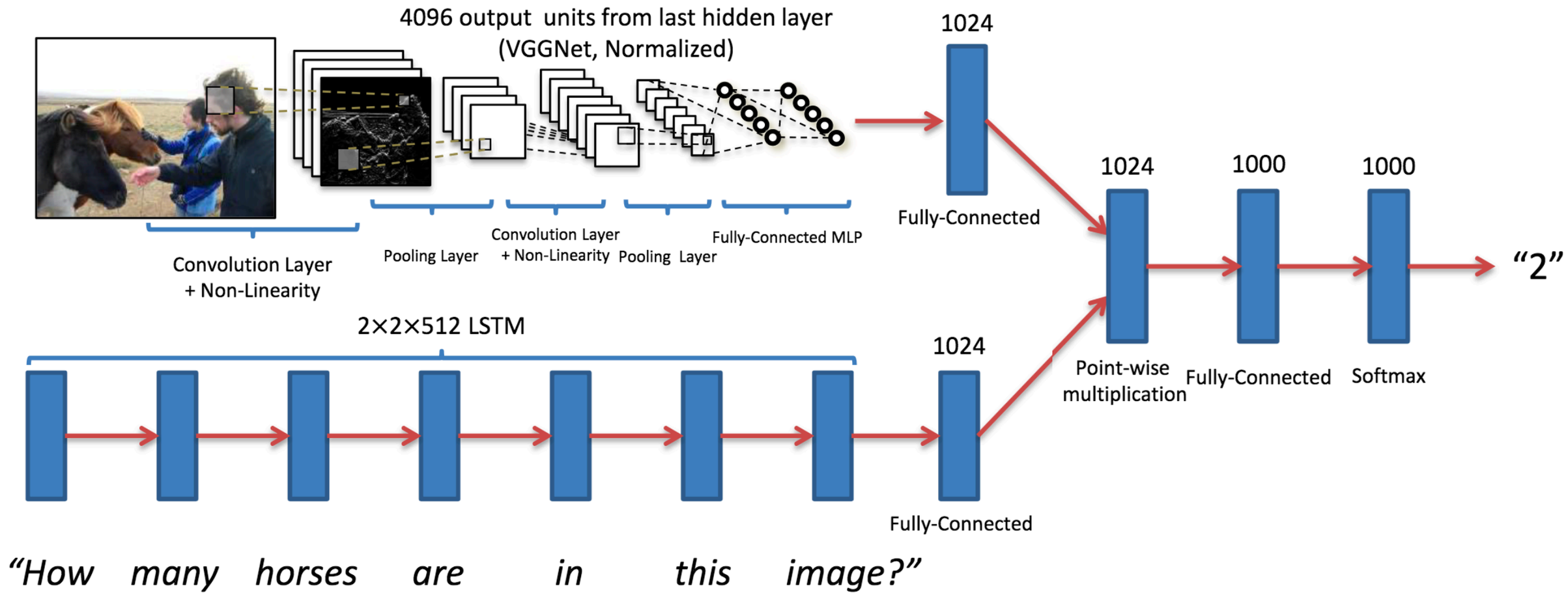
Algebraic operations with the vector representation of words

X = Vector(“Paris”) – vector(“France”) + vector(“Italy”)

Closest nearest neighbor to X is vector(“Rome”)

Architecture

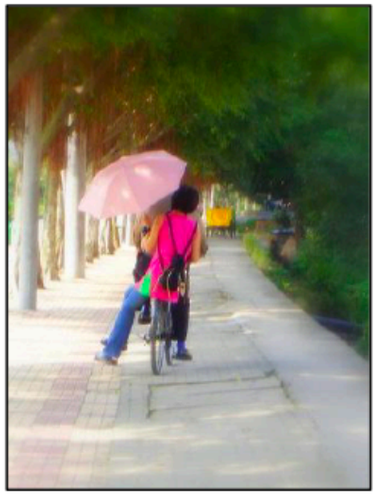
Architecture



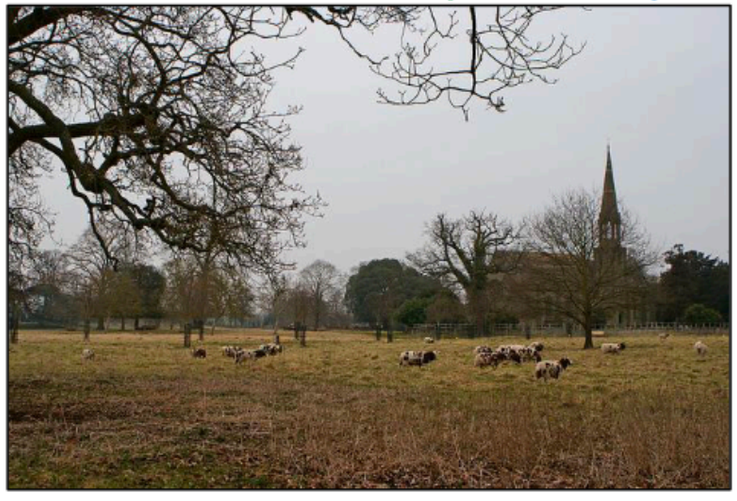
There are 1000 possible answers in this system.
Questions are unlimited.

What red objects in front are almost covered by snow?	meters parking meters parking meters	car cars shoes
Is it winter?	yes yes yes	no yes yes

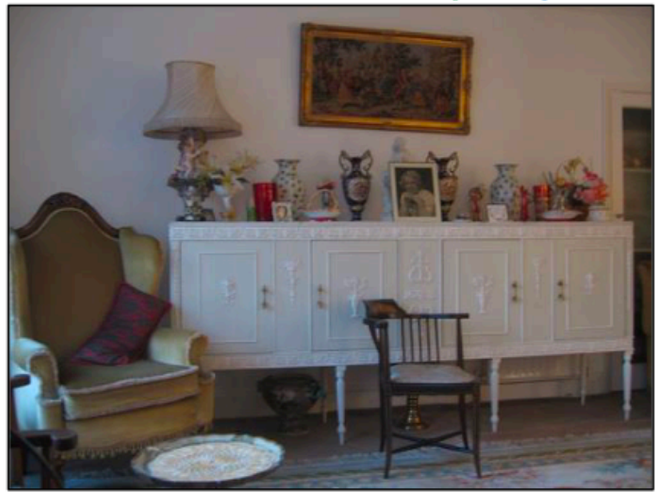
Does the car have a license plate?	yes yes yes	yes yes yes
Could the truck have a camper?	yes yes yes	yes yes yes



Is the woman on the back of the bicycle pedaling?	no no yes	no no yes
Why is the woman holding an umbrella?	sunny to block sun uncertain	it's raining it's raining to stay dry



Is this photo taken in Antarctica?	no no no	no yes yes
Overcast or sunny?	overcast overcast overcast	overcast overcast sunny



Is the picture hanging straight?	no yes yes	no yes yes
How many cabinets are on the piece of furniture?	4 4 4	3 3 6

What type of trees are here?	palm palm palm	ash oak pine
Is the skateboard airborne?	yes yes yes	no yes yes

Fig. 27: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset.

what is on the ground?

Submit

what is on the ground?

Submit

Predicted top-5 answers with confidence:

sand

90.748%

snow

2.858%

beach

1.418%

surfboards

0.677%

water

0.528%

what color is the umbrella?

Submit

what color is the umbrella?

Submit

Predicted top-5 answers with confidence:

yellow

95.090%

white

1.811%

black

0.663%

blue

0.541%

gray

0.362%

are we alone in the universe?

Submit

are we alone in the universe?

Submit

Predicted top-5 answers with confidence:

no

78.234%

yes

21.763%

people

0.001%

birds

0.000%

out

0.000%

what is the meaning of life?

Submit

what is the meaning of life?

Submit

Predicted top-5 answers with confidence:

beach

15.262%

sand

8.537%

seagull

4.708%

tower

2.393%

rocks

1.746%

what is the yellow thing?

Submit

Predicted top-5 answers with confidence:

frisbee

79.844%

surfboard

7.319%

banana

2.844%

lemon

2.438%

surfboards

1.252%

how many trains are in the picture?

Submit

Predicted top-5 answers with confidence:

3

30.233%

5

18.270%

4

17.000%

2

11.343%

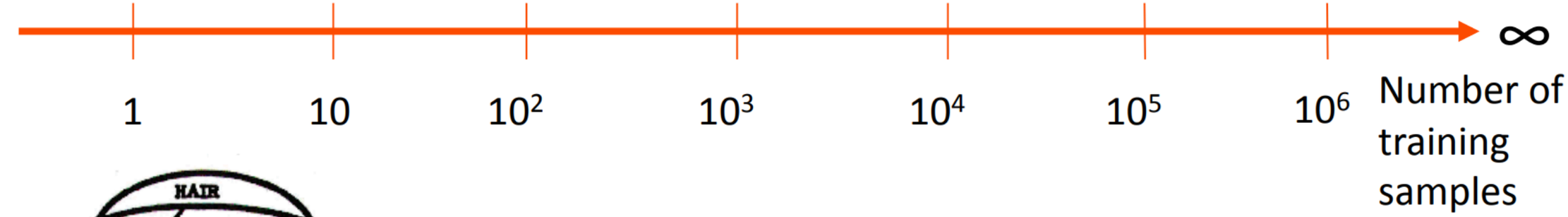
6

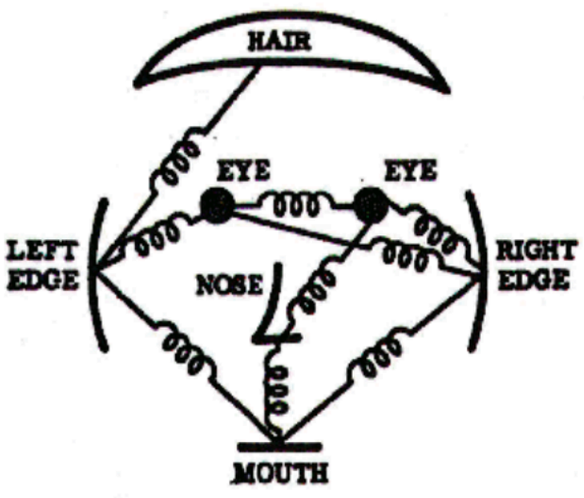
7.806%

Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences



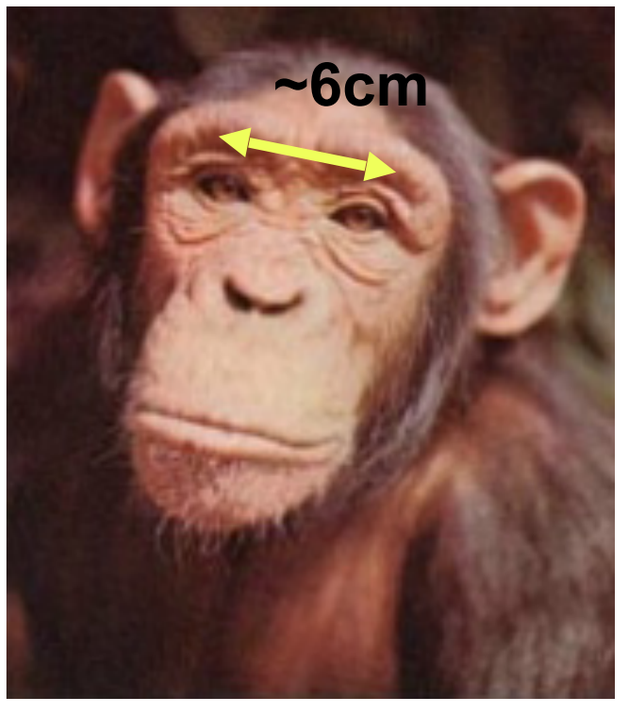


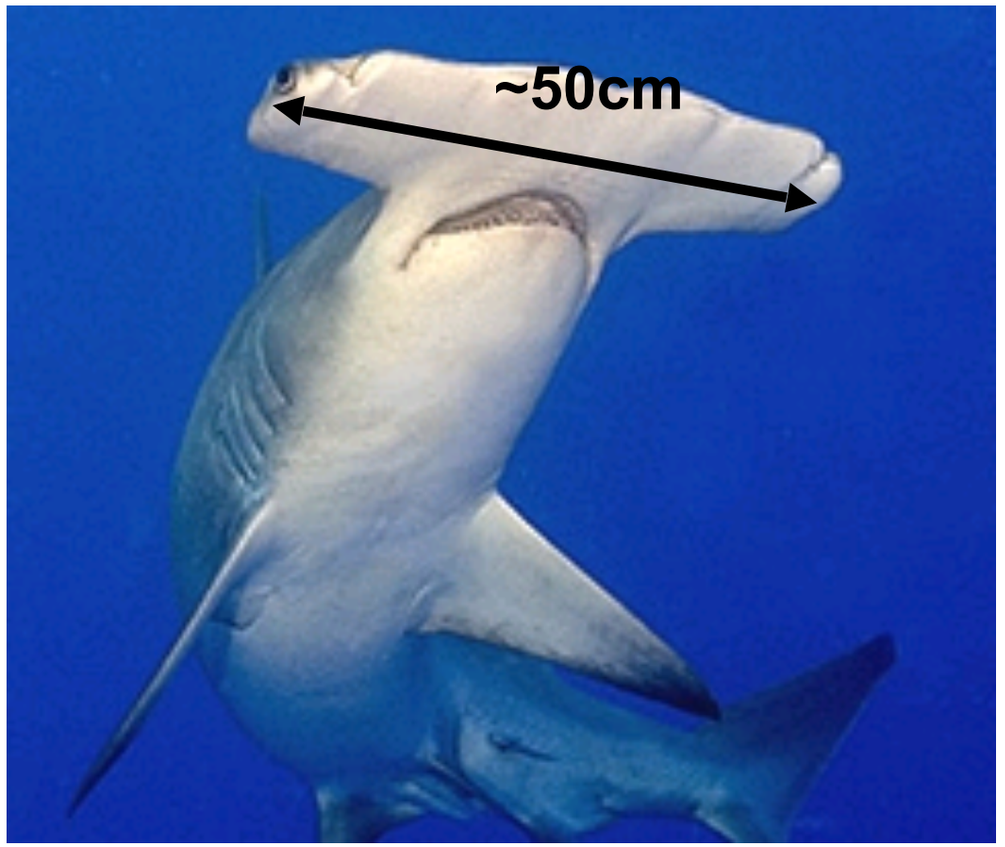
Slide credit: Aude Oliva

Stereo

Computer Vision
Fall 2019
Columbia University

Stereo vision





Slide credit: Antonio Torralba

Why not put our second eye here?

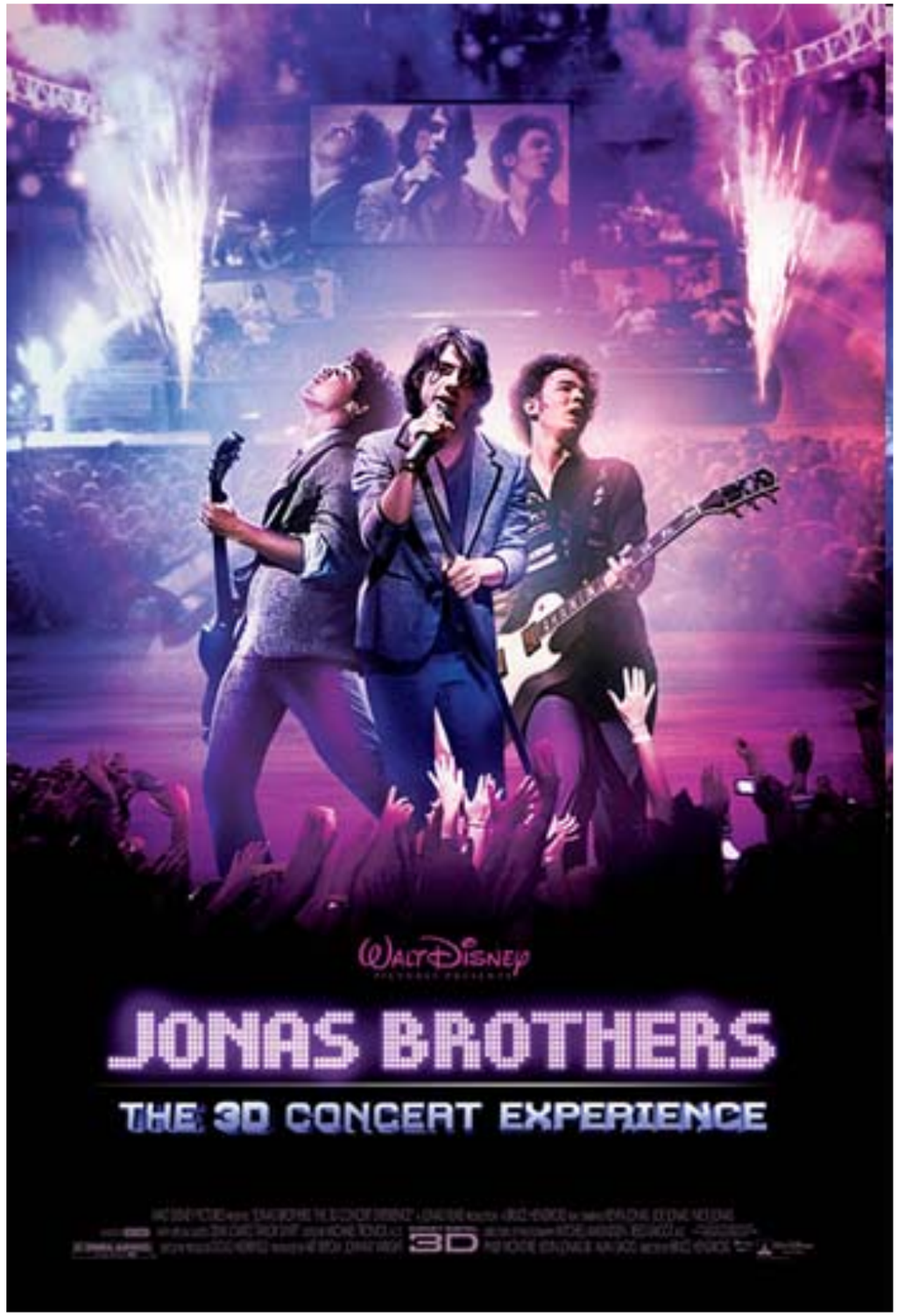
Stereoscopes: A 19th Century Pastime

Public Library, Stereoscopic Looking Room, Chicago, by Phillips, 1923

Teesta suspension bridge-Darjeeling, India

Mark Twain at Pool Table", no date, UCR Museum of Photography

3D Movies


WALT DISNEY
JONAS BROTHERS
THE 3D CONCERT EXPERIENCE
3D


MY BLOODY VALENTINE
3D
GET YOUR HEART BROKEN
IN THEATERS JANUARY 19

Depth without objects

Random dot stereograms (Bela Julesz)

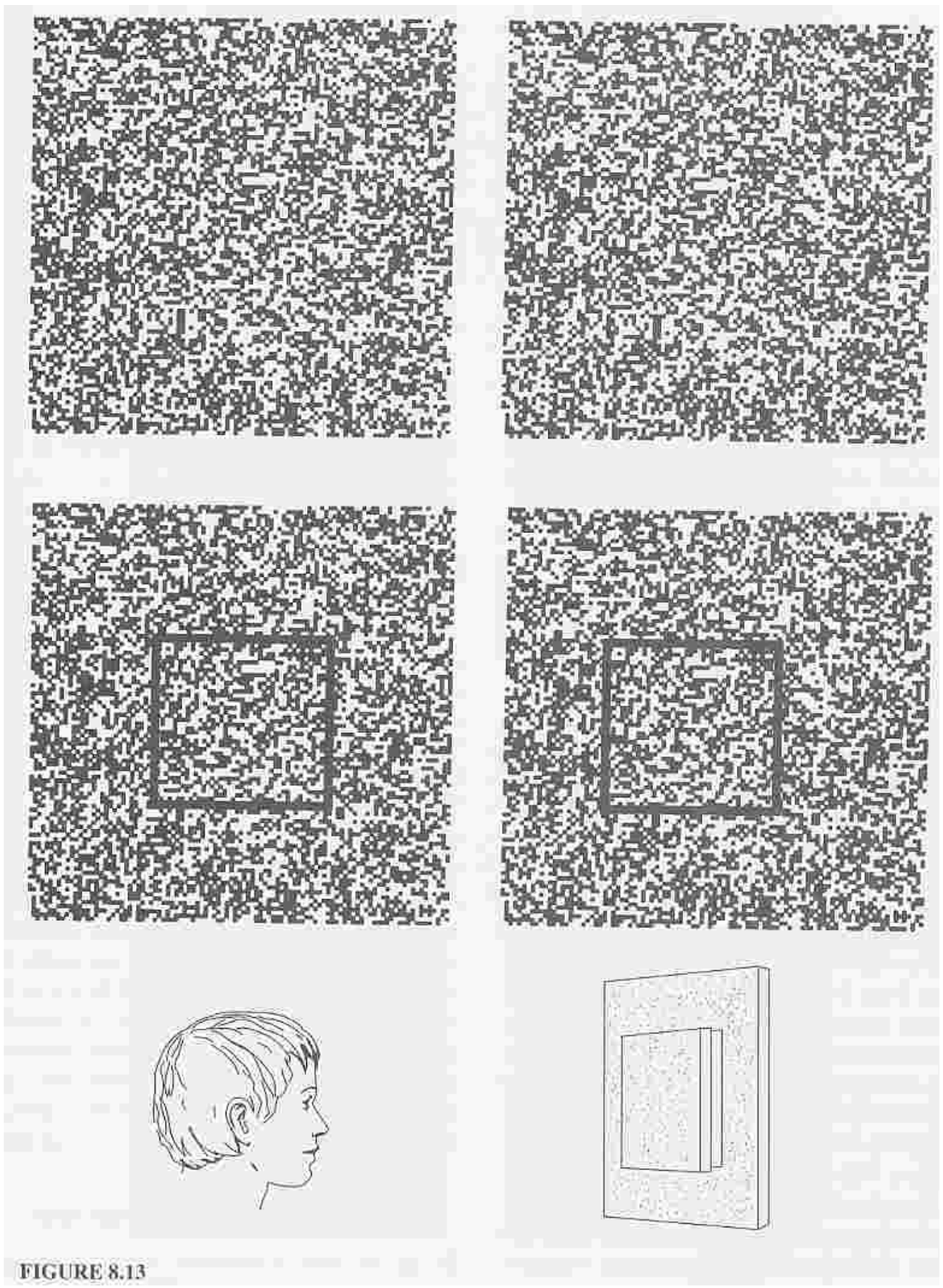
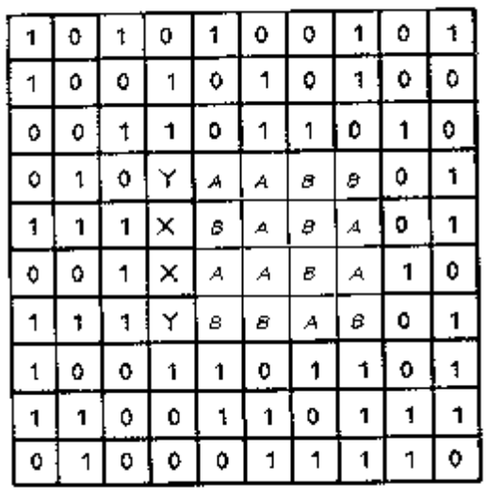
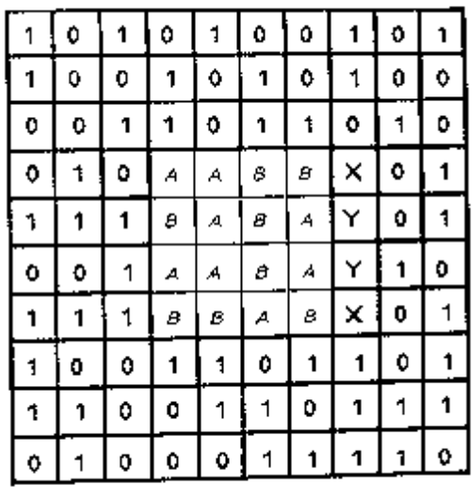


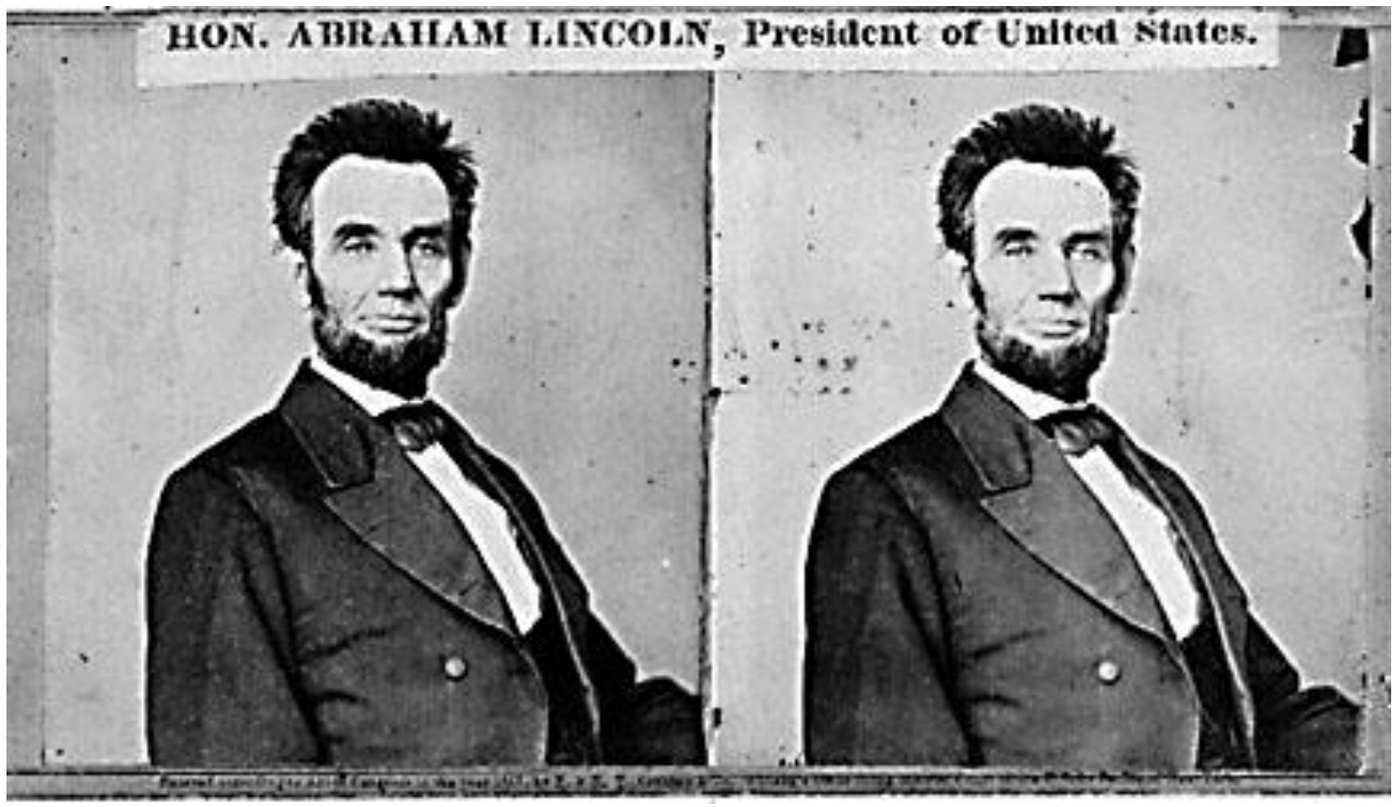
FIGURE 8.13





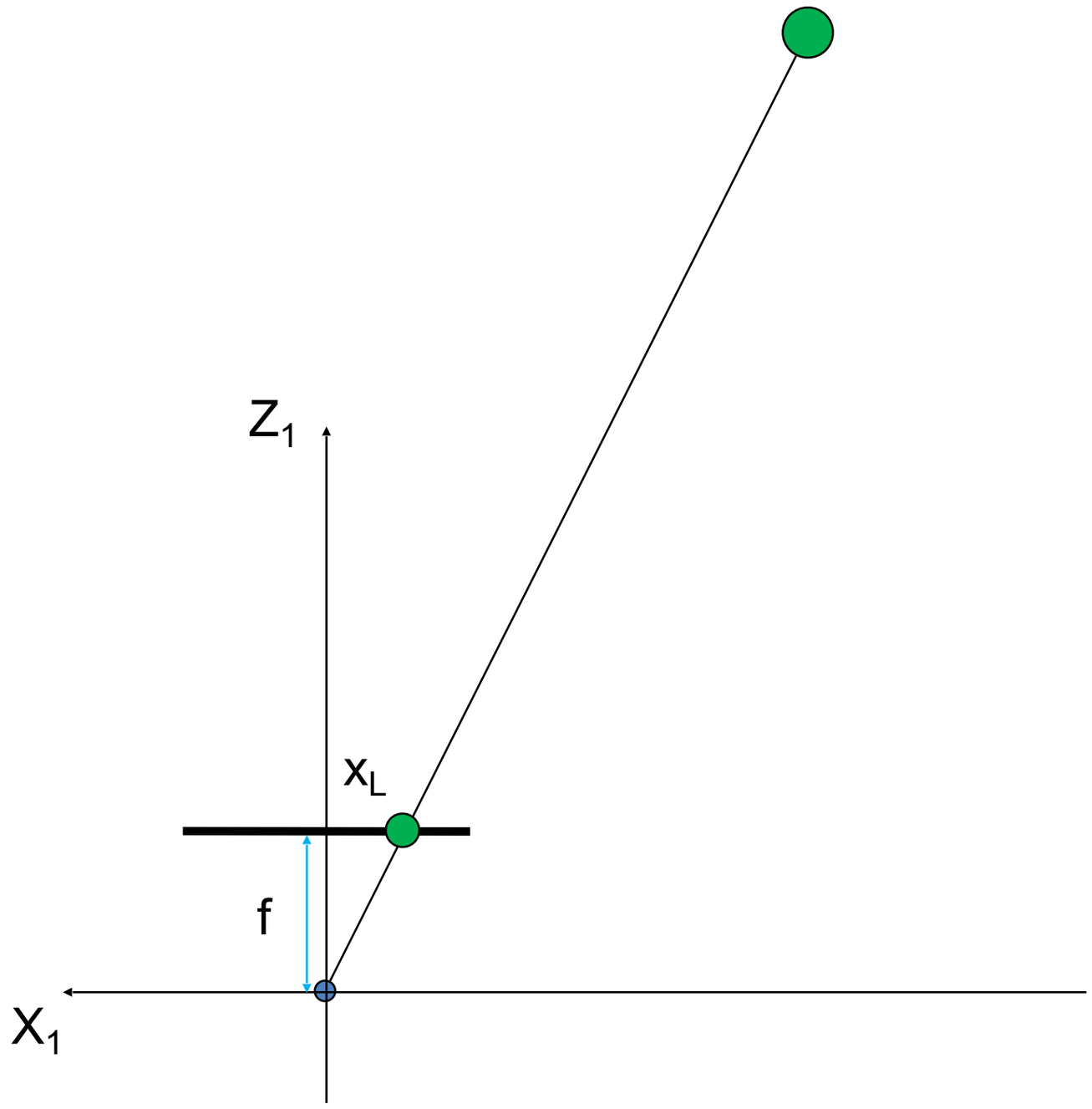
Julesz, 1971

Stereo



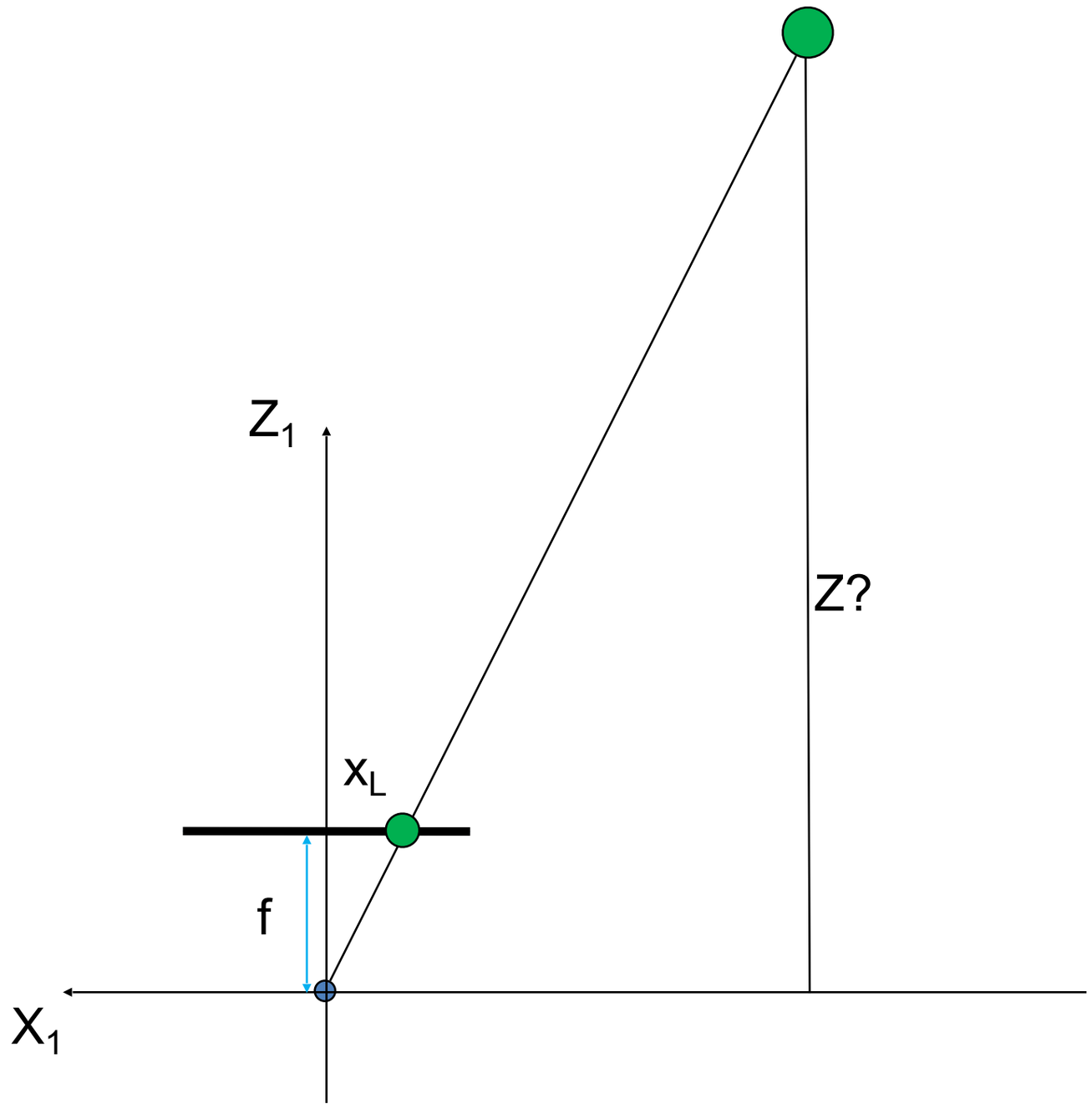
- Given two images from different viewpoints
  - How can we compute the depth of each point in the image?
  - Based on *how much each pixel moves* between the two images

Geometry for a simple stereo system



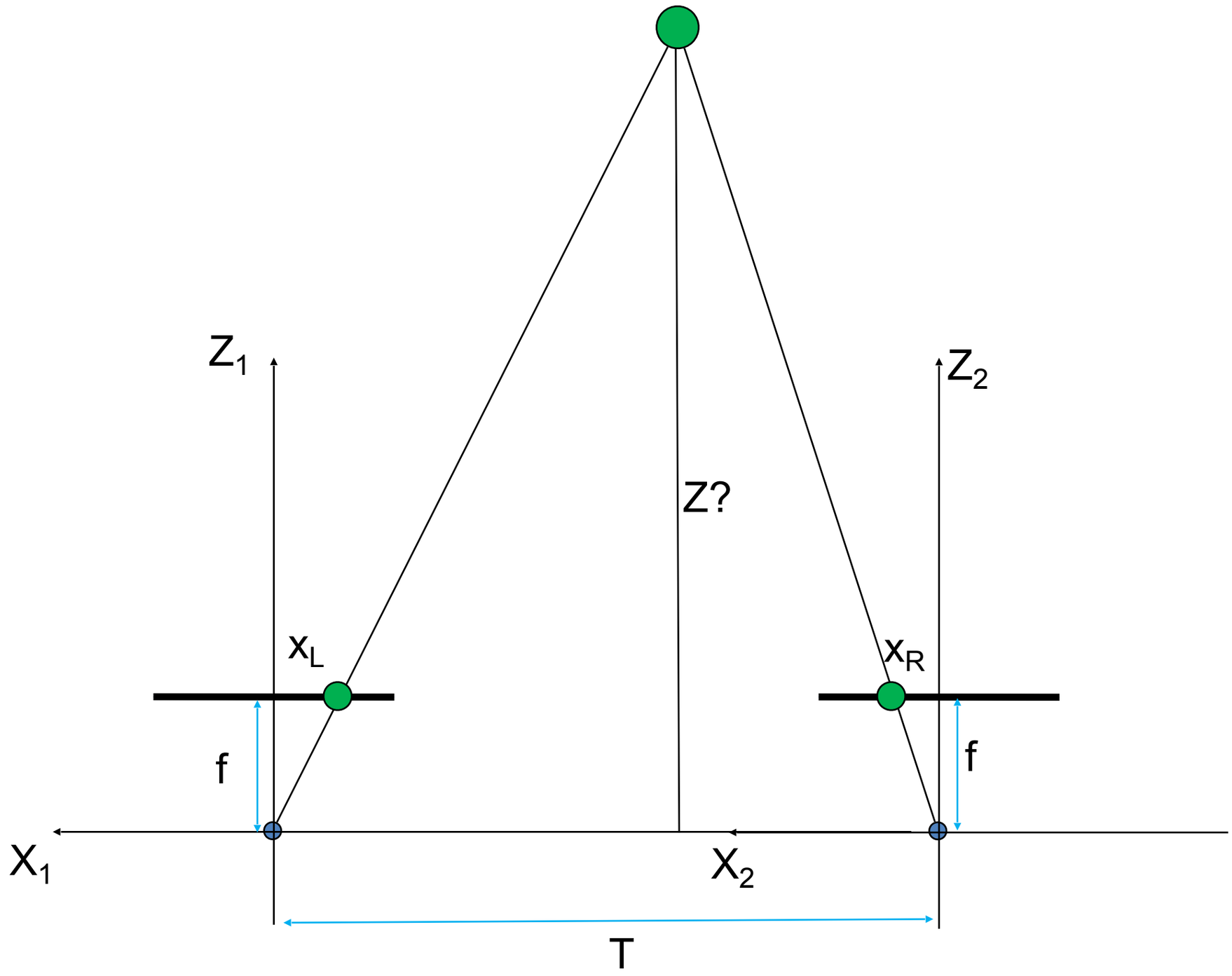
Slide credit: Antonio Torralba

Geometry for a simple stereo system



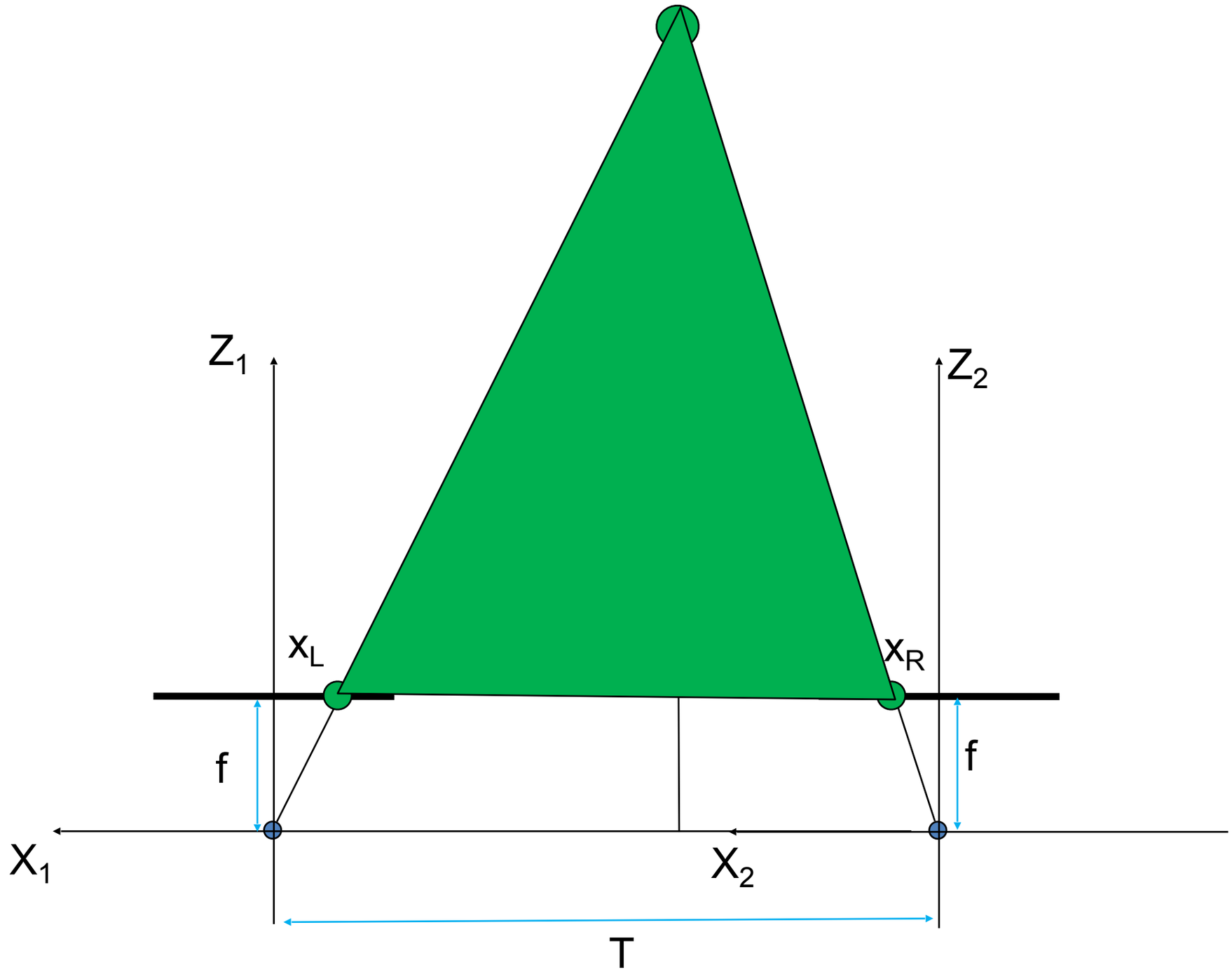
Slide credit: Antonio Torralba

Geometry for a simple stereo system



Slide credit: Antonio Torralba

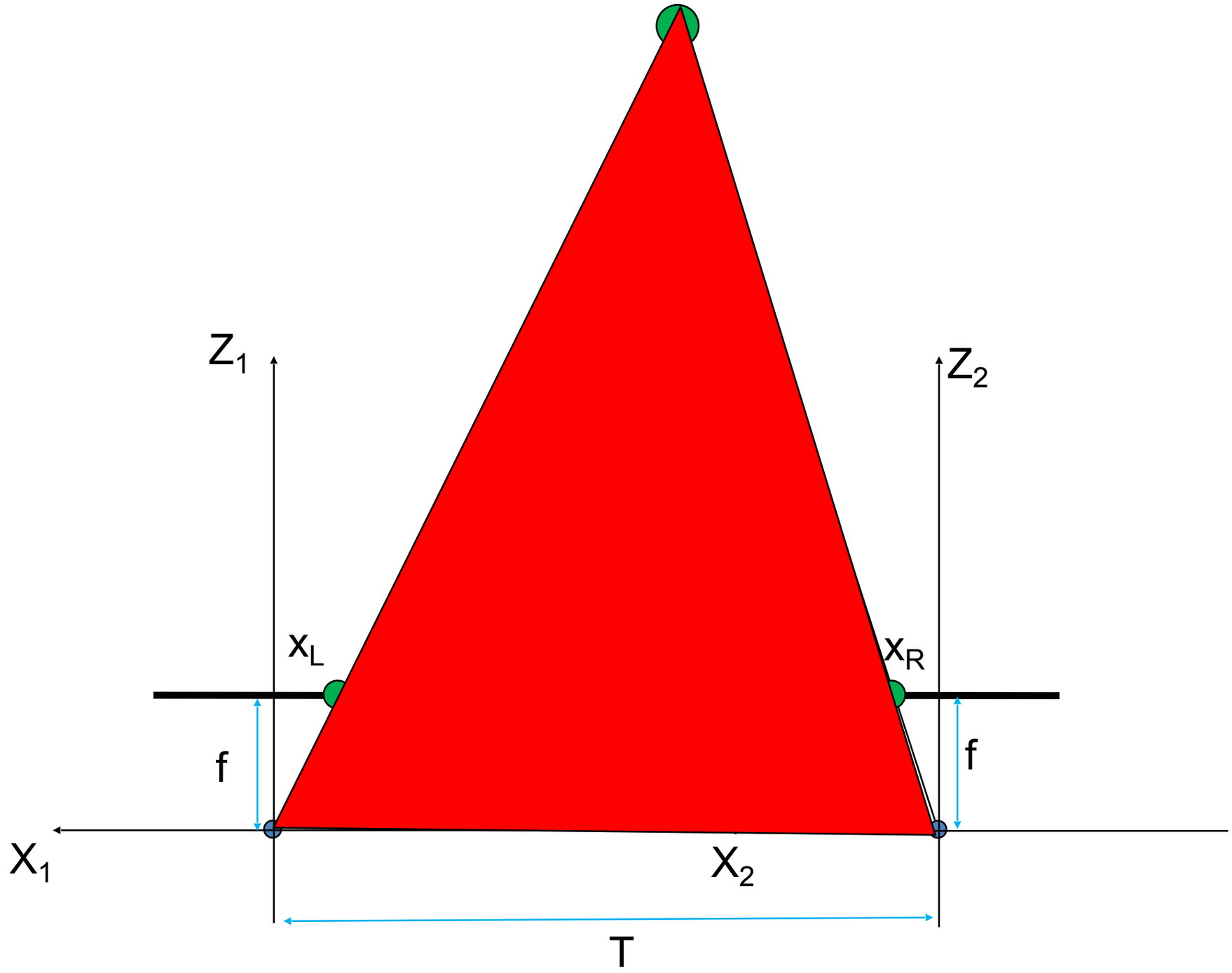
Geometry for a simple stereo system



Similar triangles

Slide credit: Antonio Torralba

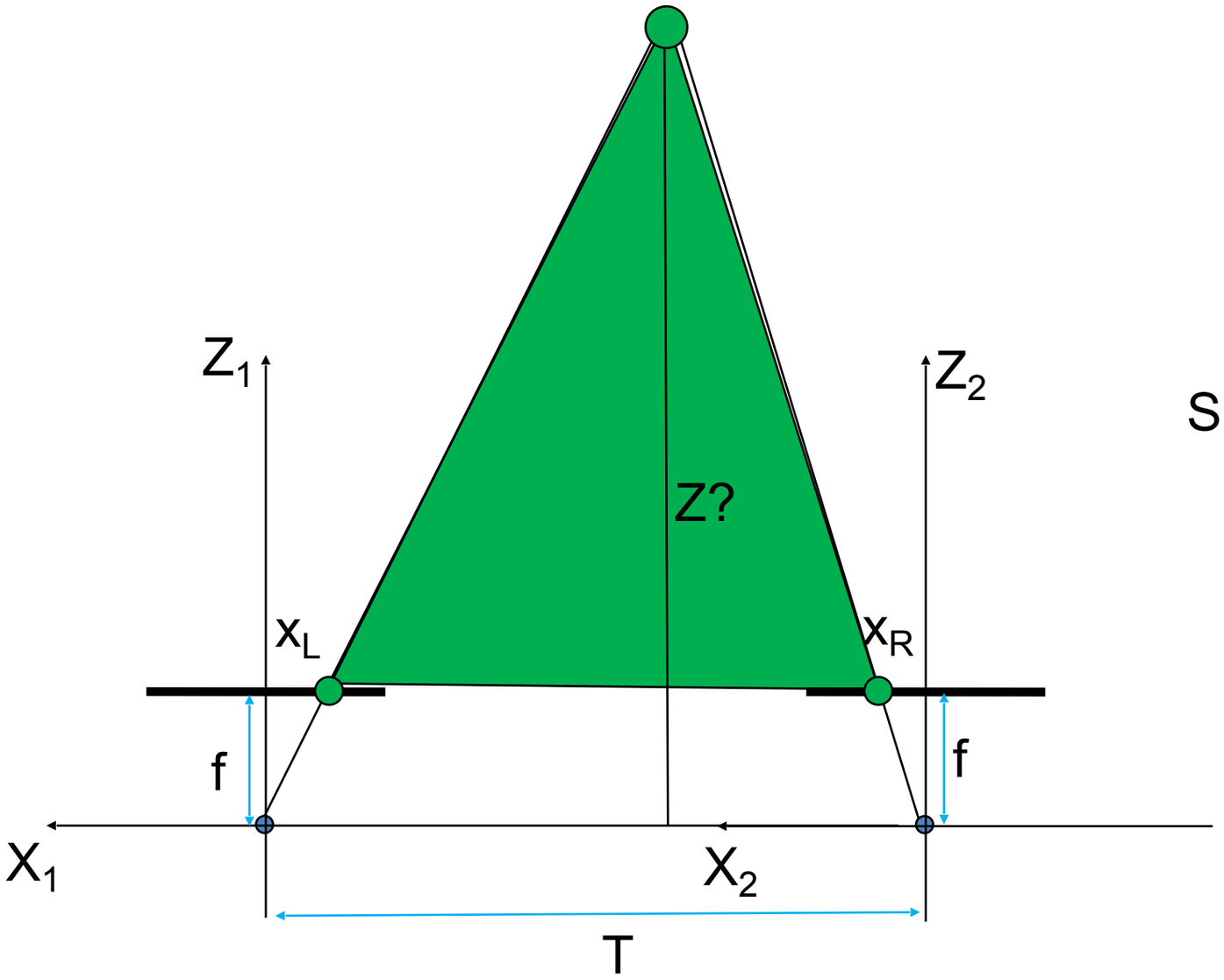
Geometry for a simple stereo system



Similar triangles

Slide credit: Antonio Torralba

Geometry for a simple stereo system

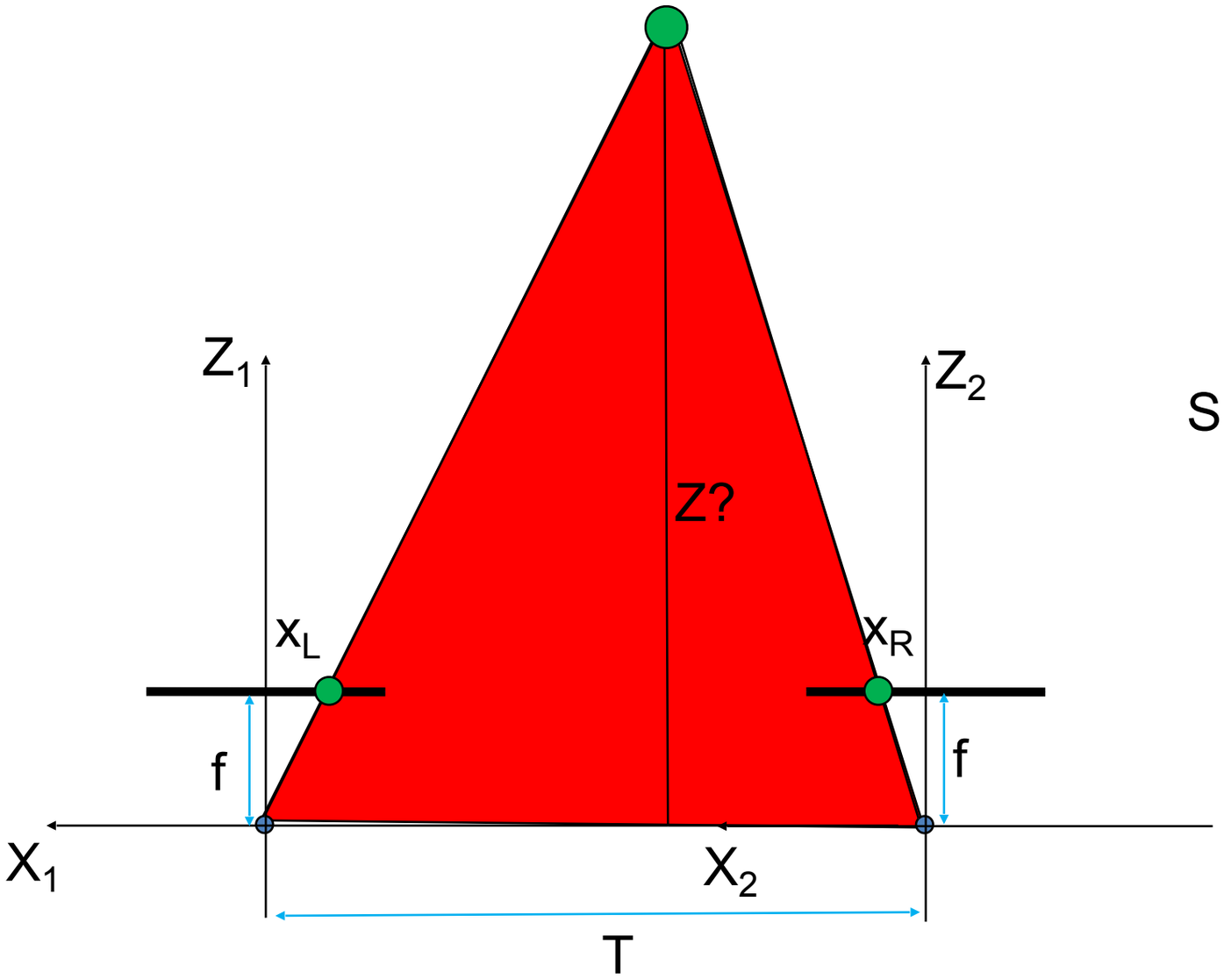


Similar triangles:

$$\frac{T+X_L-X_R}{Z-f} =$$

Slide credit: Antonio Torralba

Geometry for a simple stereo system

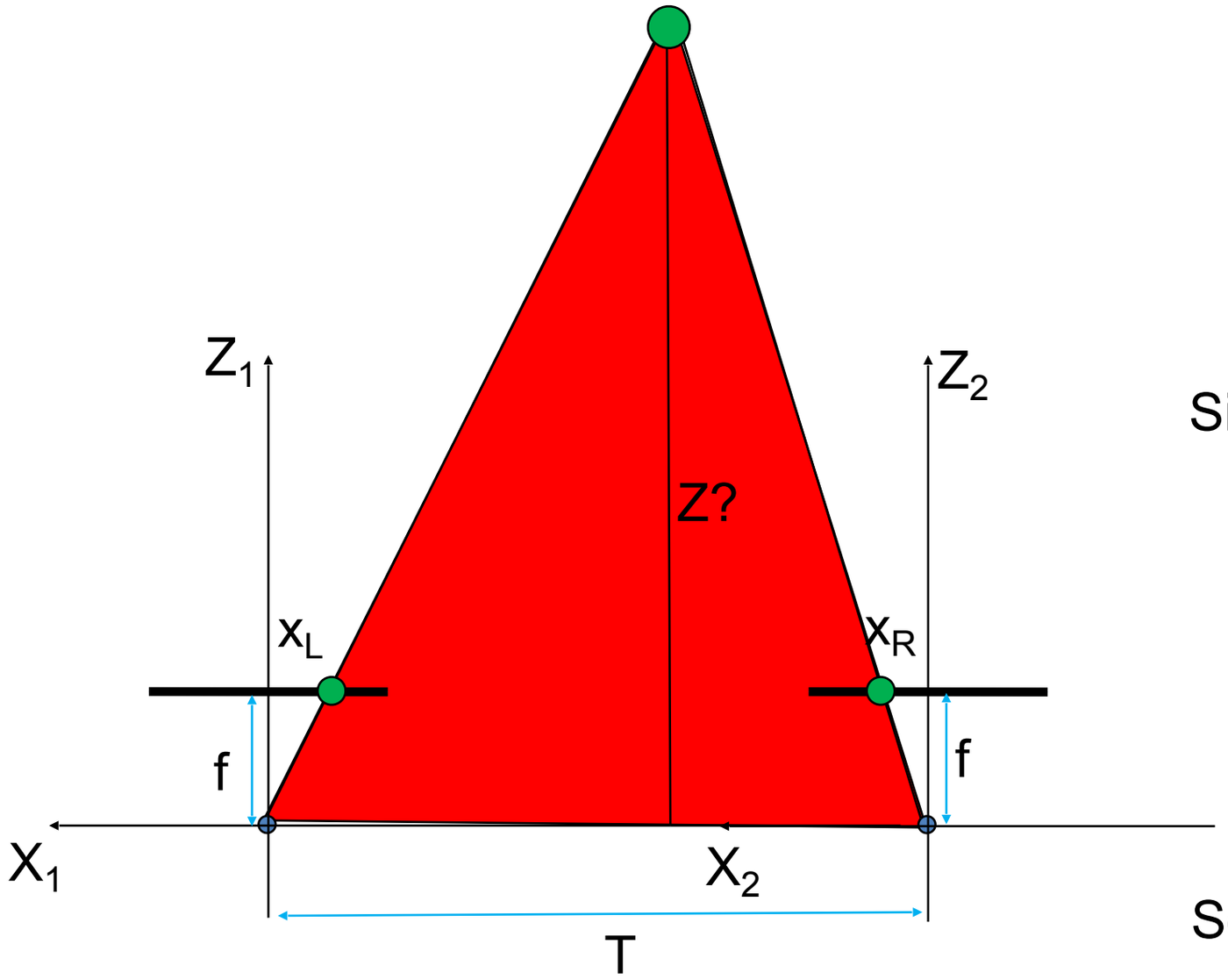


Similar triangles:

$$\frac{T+X_L-X_R}{Z-f} = \frac{T}{Z}$$

Slide credit: Antonio Torralba

Geometry for a simple stereo system



Similar triangles:

$$\frac{T+X_L-X_R}{Z-f} = \frac{T}{Z}$$

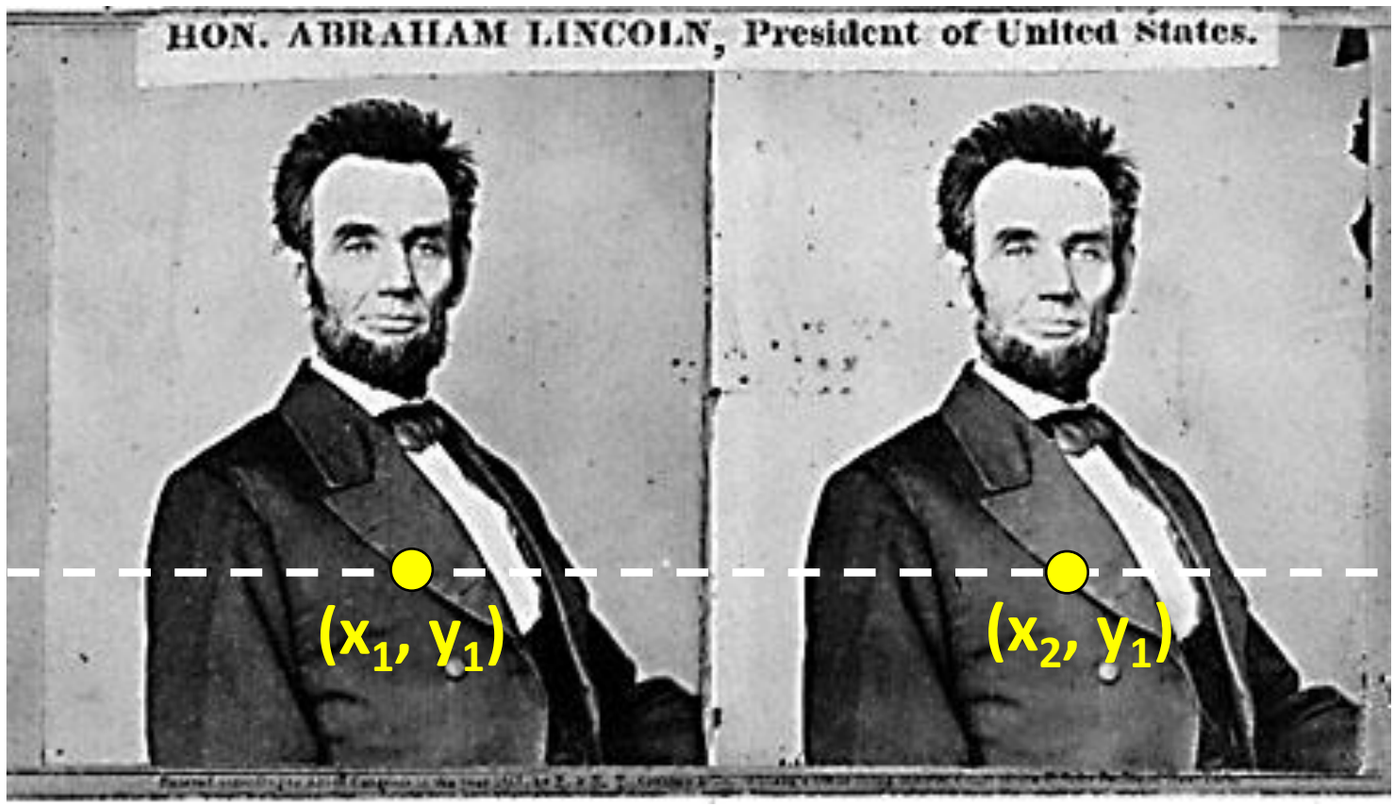
Solving for Z:

$$Z = f\,\frac{T}{X_R - X_L}$$

Disparity

Slide credit: Antonio Torralba

Epipolar geometry

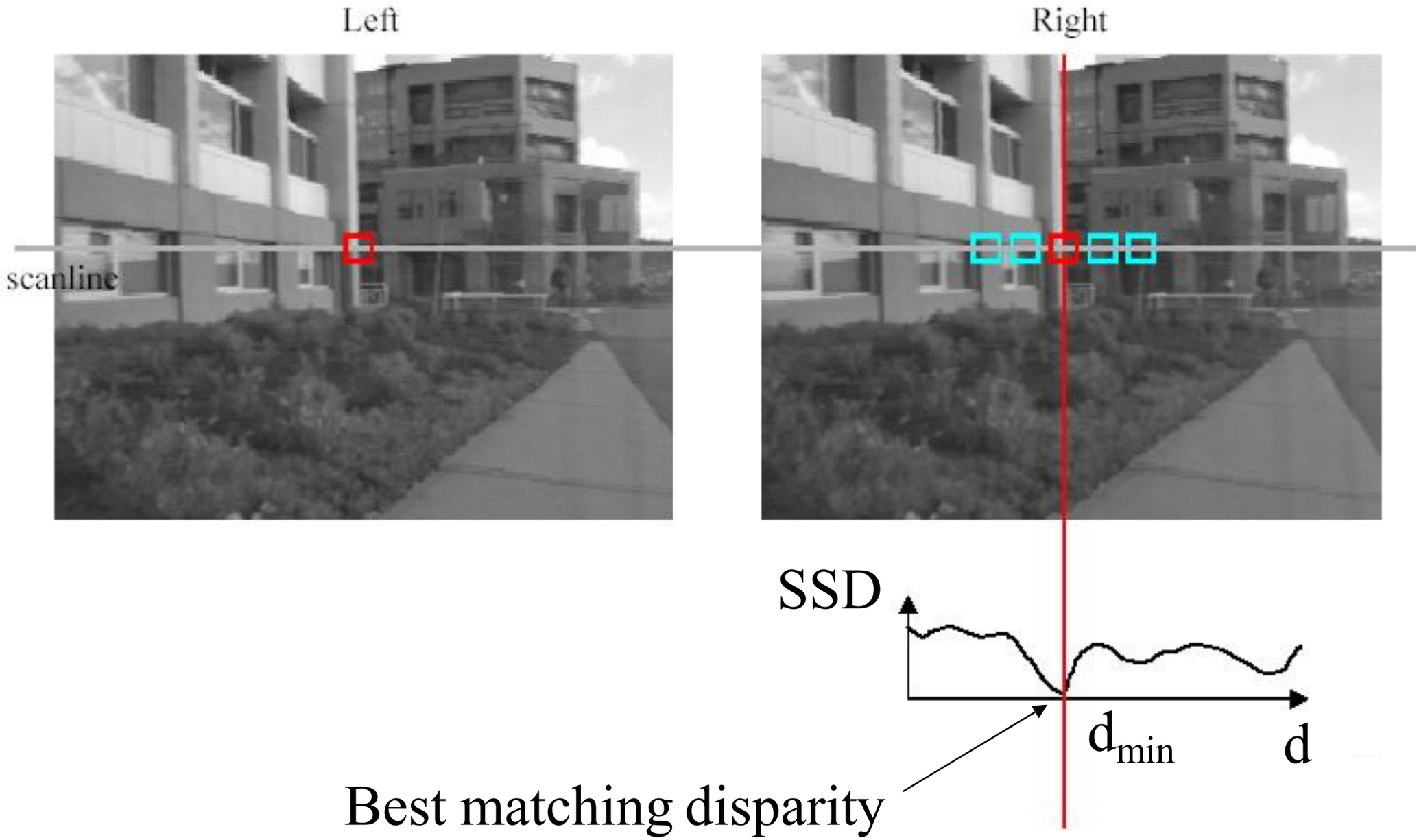


Two images captured by a purely horizontal translating camera (*rectified* stereo pair)

$x_2 - x_1$ = the *disparity* of pixel (x_1, y_1)

Slide credit: Noah Snavely

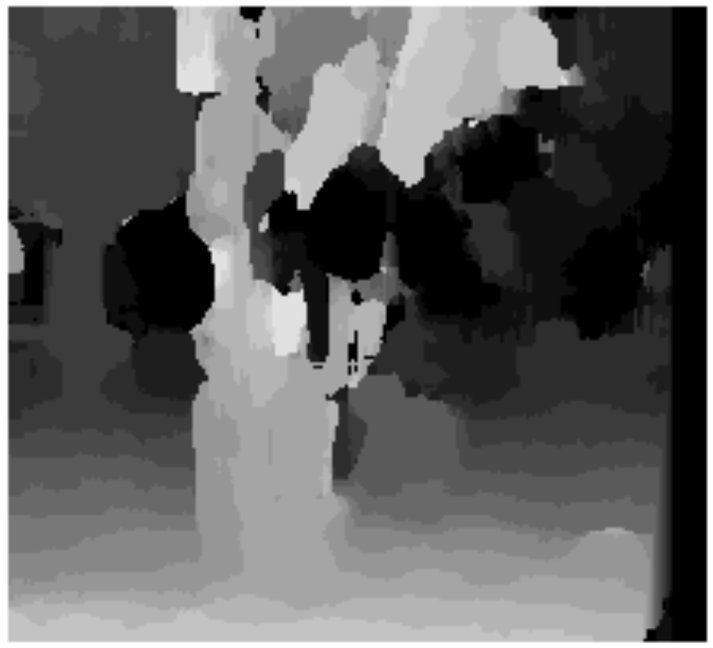
Stereo matching based on SSD



Slide credit: Noah Snavely

Window size

$W = 3$



$W = 20$

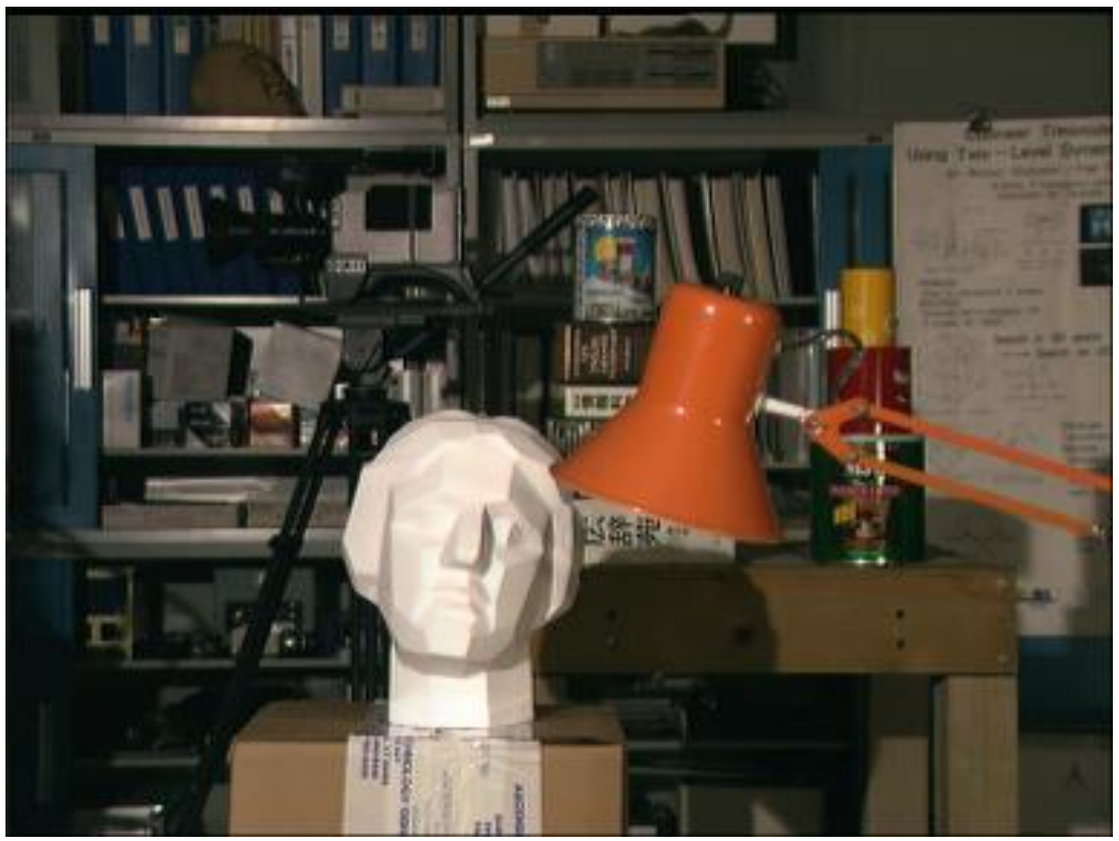
Effect of window size

Better results with *adaptive window*

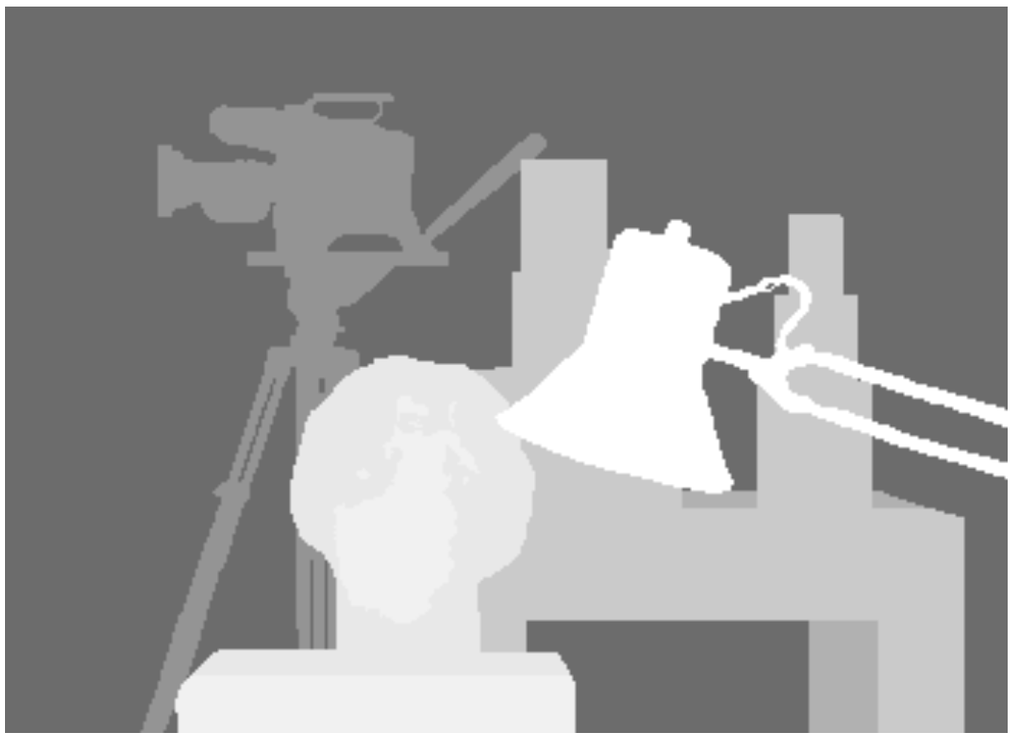
- T. Kanade and M. Okutomi, *A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment*,, Proc. International Conference on Robotics and Automation, 1991.
- D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. International Journal of Computer Vision, 28(2):155-174, July 1998

Slide credit: Noah Snavely

Middlebury Dataset



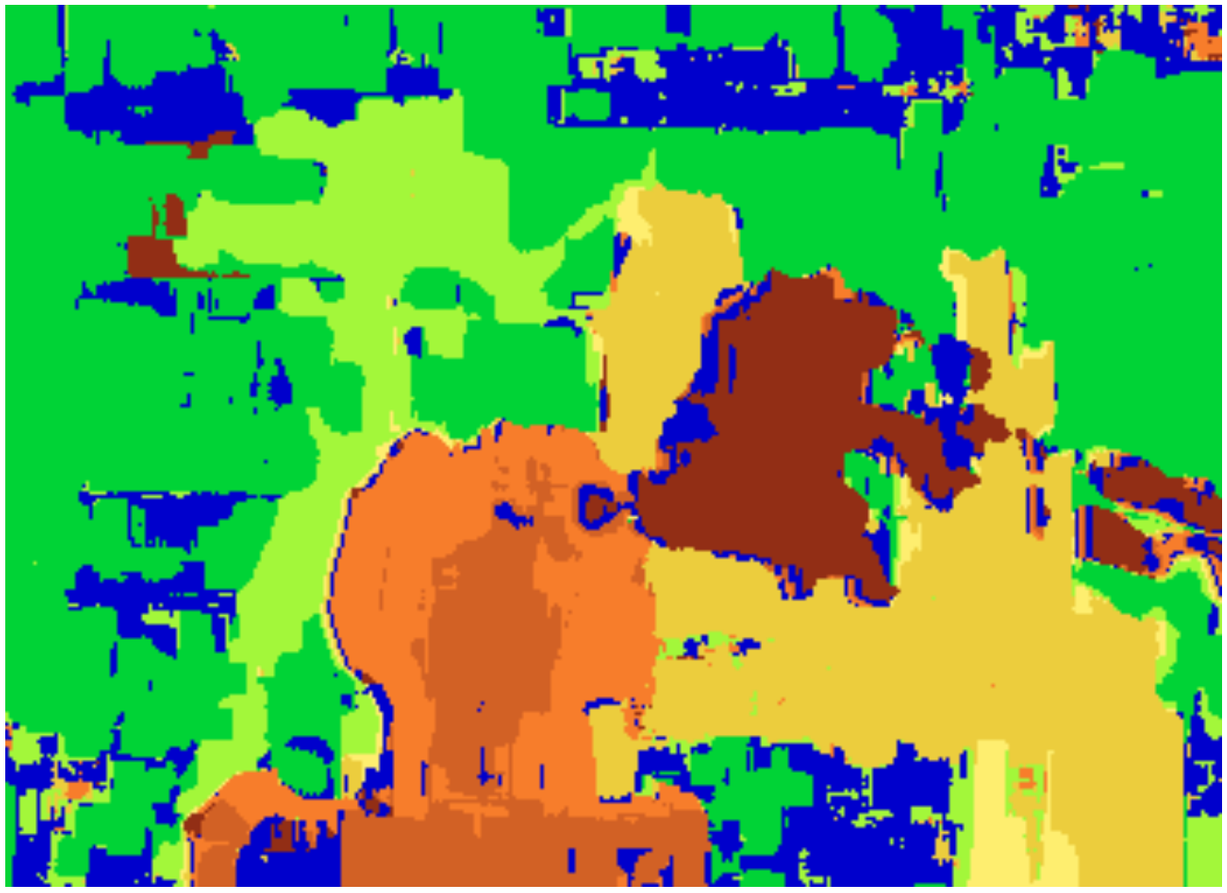
Scene



Ground truth

Slide credit: Noah Snavely

Results with window search

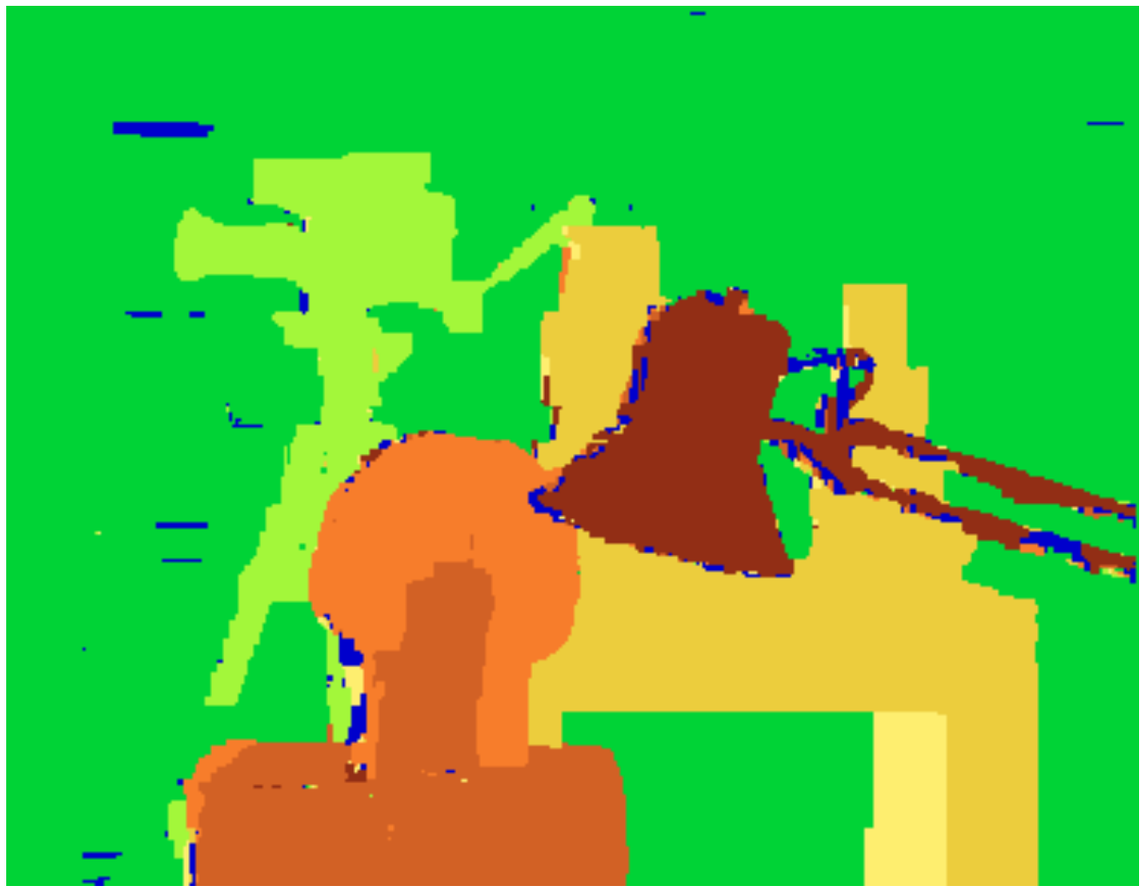


Window-based matching
(best window size)

Ground truth

Slide credit: Noah Snavely

Better methods exist...



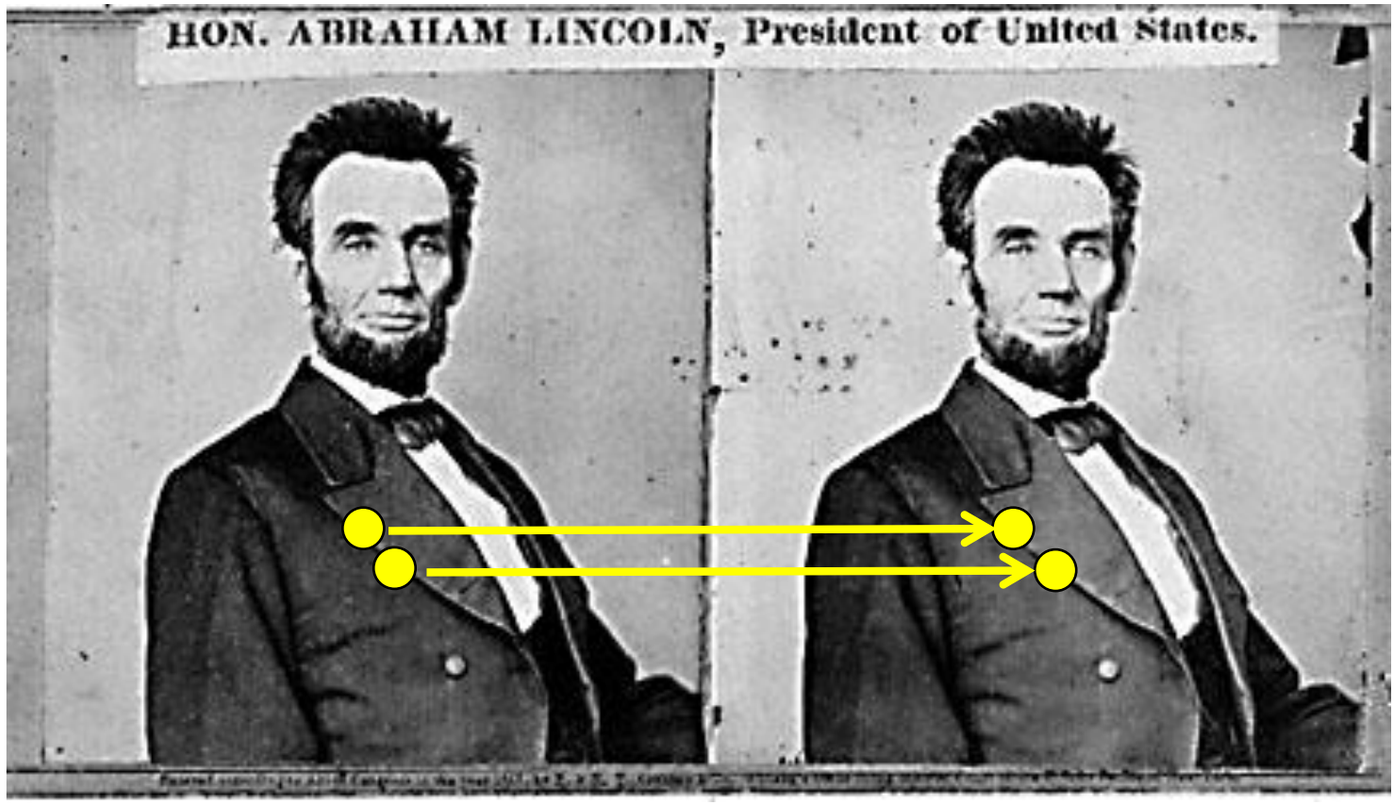
State of the art method

Boykov et al., Fast Approximate Energy Minimization via Graph Cuts,
International Conference on Computer Vision, September 1999.

Ground truth

For the latest and greatest: http://www.middlebury.edu/stereo/

Stereo as energy minimization



- What defines a good stereo correspondence?
  1. Match quality
     - Want each pixel to find a good match in the other image
  2. Smoothness
     - If two pixels are adjacent, they should (usually) move about the same amount

Slide credit: Noah Snavely

Stereo as energy minimization

- Find disparity map *d* that minimizes an energy function $E(d)$

- Simple pixel / window matching

$$E(d) = \sum_{(x,y) \in I} C(x, y, d(x, y))$$

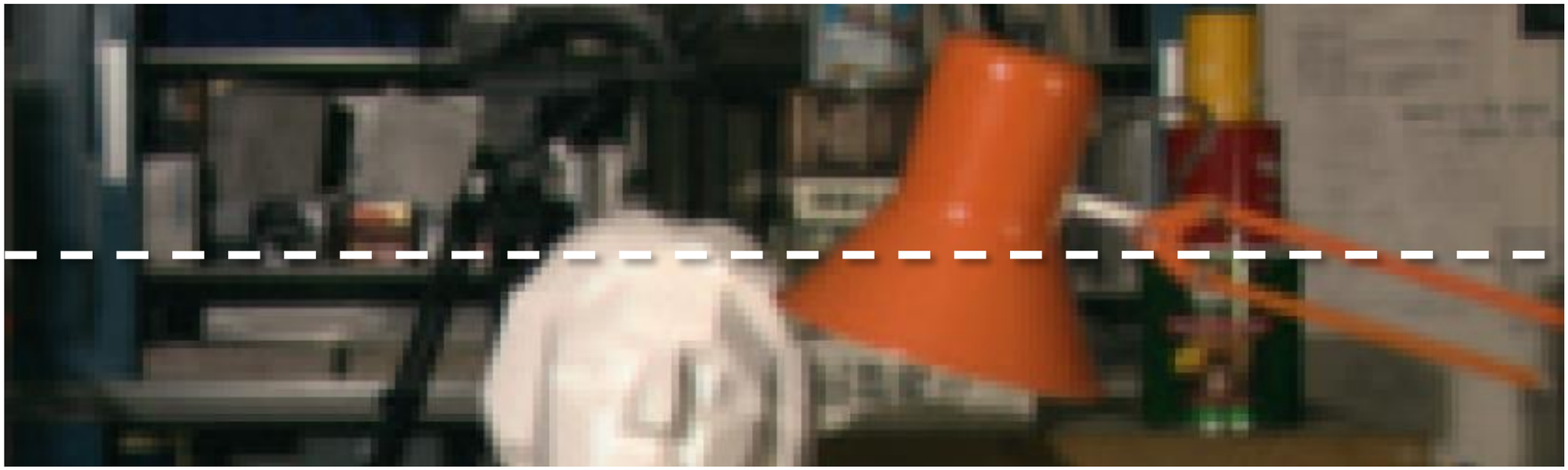
$C(x, y, d(x, y))$ = SSD distance between windows $I(x, y)$ and $J(x + d(x,y), y)$

Slide credit: Noah Snavely

Stereo as energy minimization

$I(x, y)$

$J(x, y)$



$C(x, y, d)$; the *disparity space image* (DSI)

Slide credit: Noah Snavely

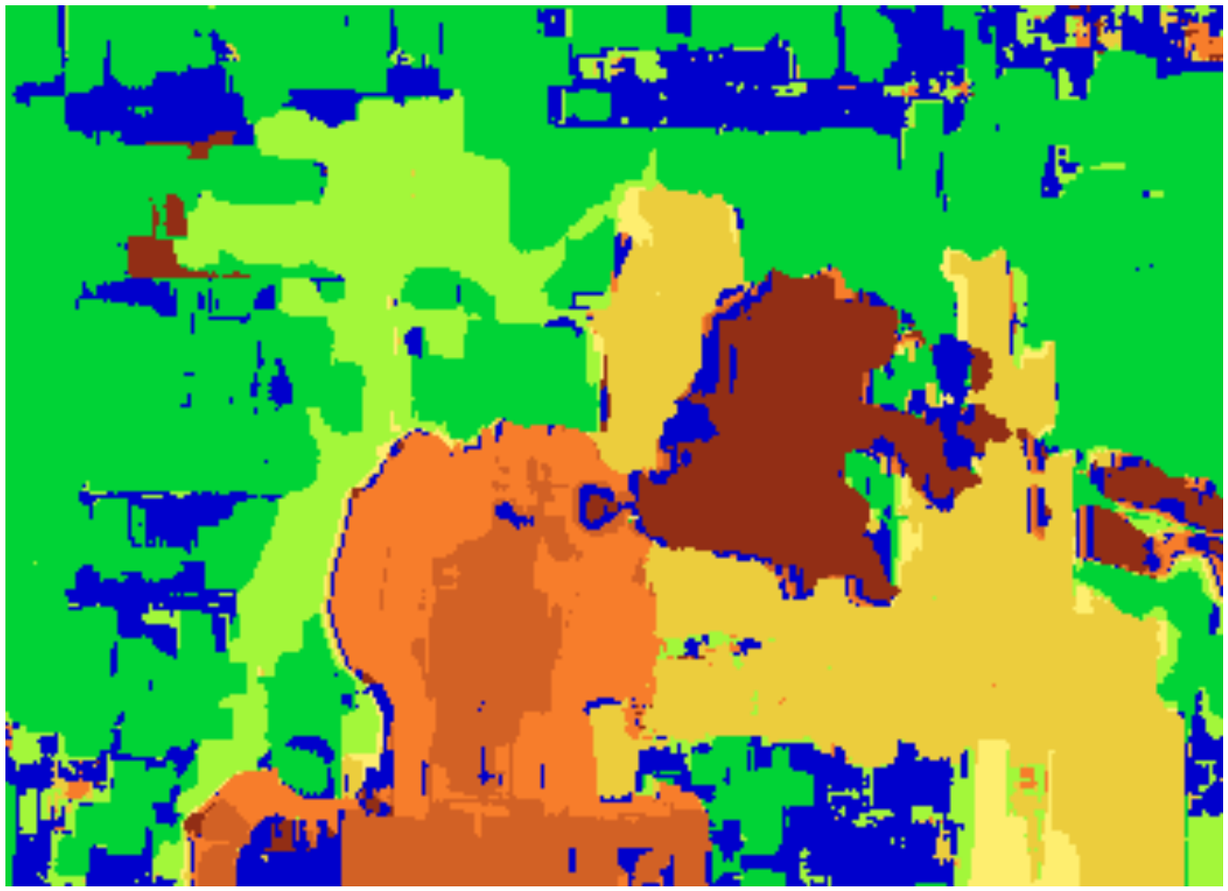
Stereo as energy minimization

Simple pixel / window matching: choose the minimum of each column in the DSI independently:

$$d(x,y) = \arg\min_{d'} C(x,y,d')$$

Slide credit: Noah Snavely

Greedy selection of best match



Slide credit: Noah Snavely

Stereo as energy minimization

- Better objective function

$$E(d) = \underbrace{E_d(d)}_{\text{match cost}} + \lambda \underbrace{E_s(d)}_{\text{smoothness cost}}$$

Want each pixel to find a good match in the other image

Adjacent pixels should (usually) move about the same amount

Slide credit: Noah Snavely

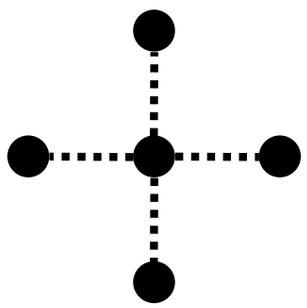
Stereo as energy minimization

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

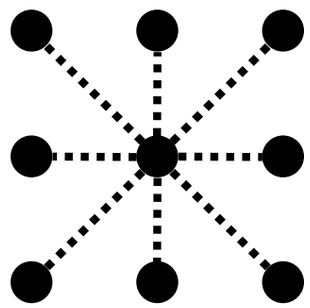
match cost: $$E_d(d) = \sum_{(x,y)\in I} C(x, y, d(x, y))$$

smoothness cost: $$E_s(d) = \sum_{(p,q)\in\mathcal{E}} V(d_p, d_q)$$

$\mathcal{E}$: set of neighboring pixels



4-connected neighborhood



8-connected neighborhood

Slide credit: Noah Snavely

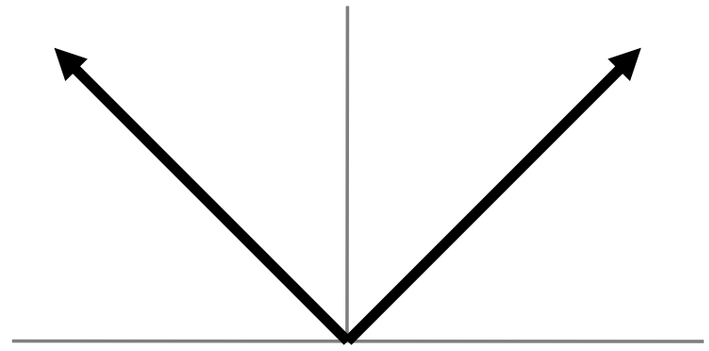
Smoothness cost

$$E_s(d) = \sum_{(p,q)\in\mathcal{E}} V(d_p, d_q)$$

How do we choose *V*?

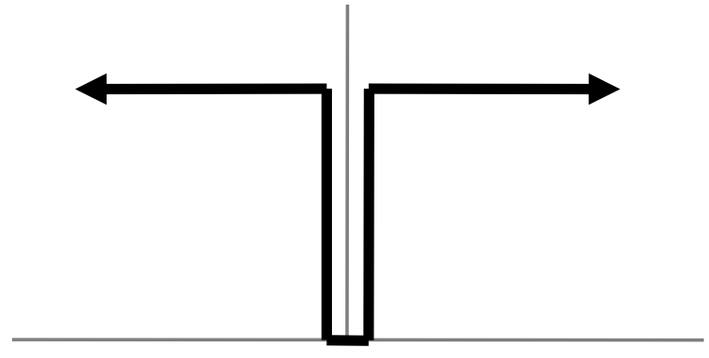
$$V(d_p, d_q) = |d_p - d_q|$$

L_1 distance



$$V(d_p, d_q) = \begin{cases} 0 & \text{if } d_p = d_q \\ 1 & \text{if } d_p \neq d_q \end{cases}$$

“Potts model”



Slide credit: Noah Snavely

Dynamic programming

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

- Can minimize this independently per scanline using dynamic programming (DP)

- Basic idea: incrementally build a table of costs *D* one column at a time

$D(x, y, i)$: minimum cost of solution such that $d(x,y) = i$

Base case: $D(0, y, i) = C(0, y, i), i = 0, \ldots, L$ (L = max disparity)

Recurrence: $D(x, y, i) = C(x, y, i) + \min_{j \in \{0,1,\ldots,L\}} D(x-1, y, j) + \lambda|i - j|$

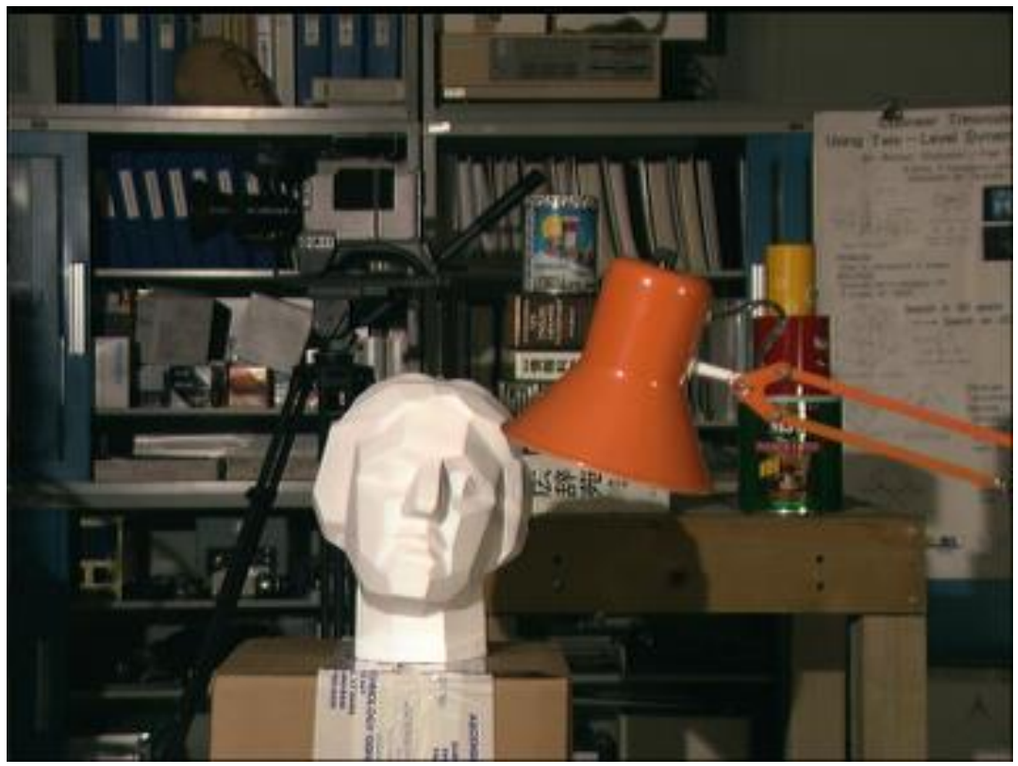
Slide credit: Noah Snavely

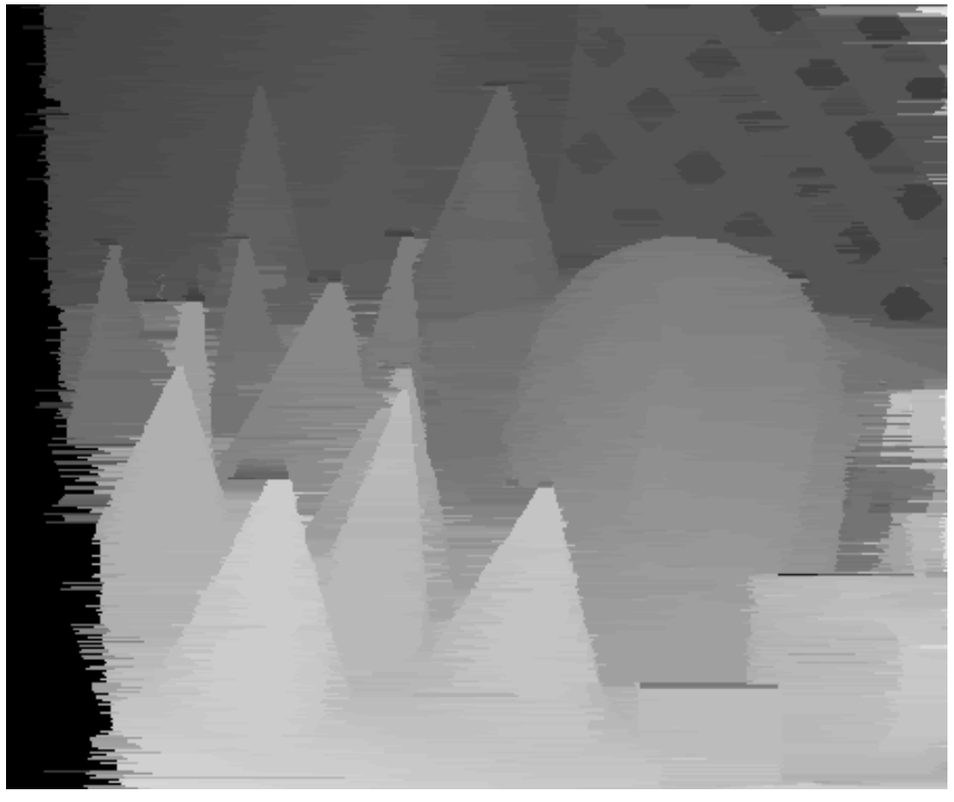
Dynamic programming

- Finds “smooth”, low-cost path through DPI from left to right

Slide credit: Noah Snavely

Dynamic Programming





Stereo as a minimization problem

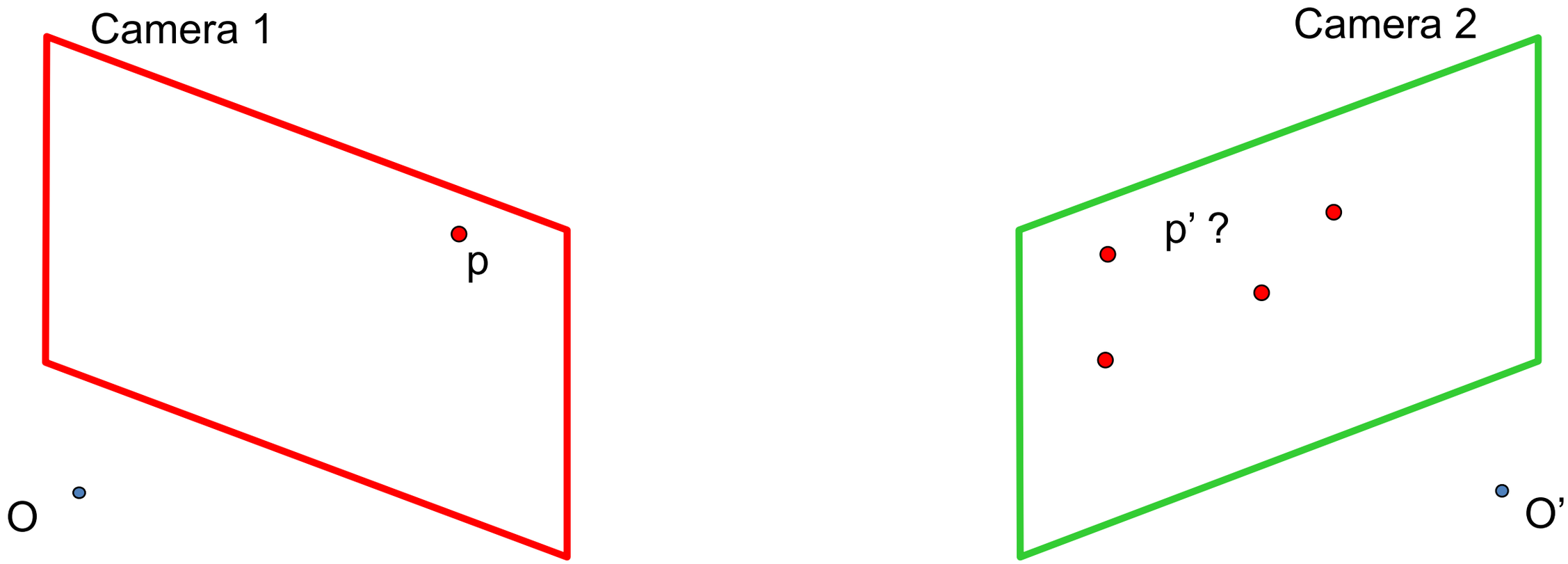
$$E(d) = E_d(d) + \lambda E_s(d)$$

- The 2D problem has many local minima
  - Gradient descent doesn't work well

- And a large search space
  - *n* x *m* image w/ *k* disparities has k^{nm} possible solutions
  - Finding the global minimum is NP-hard in general

Slide credit: Noah Snavely

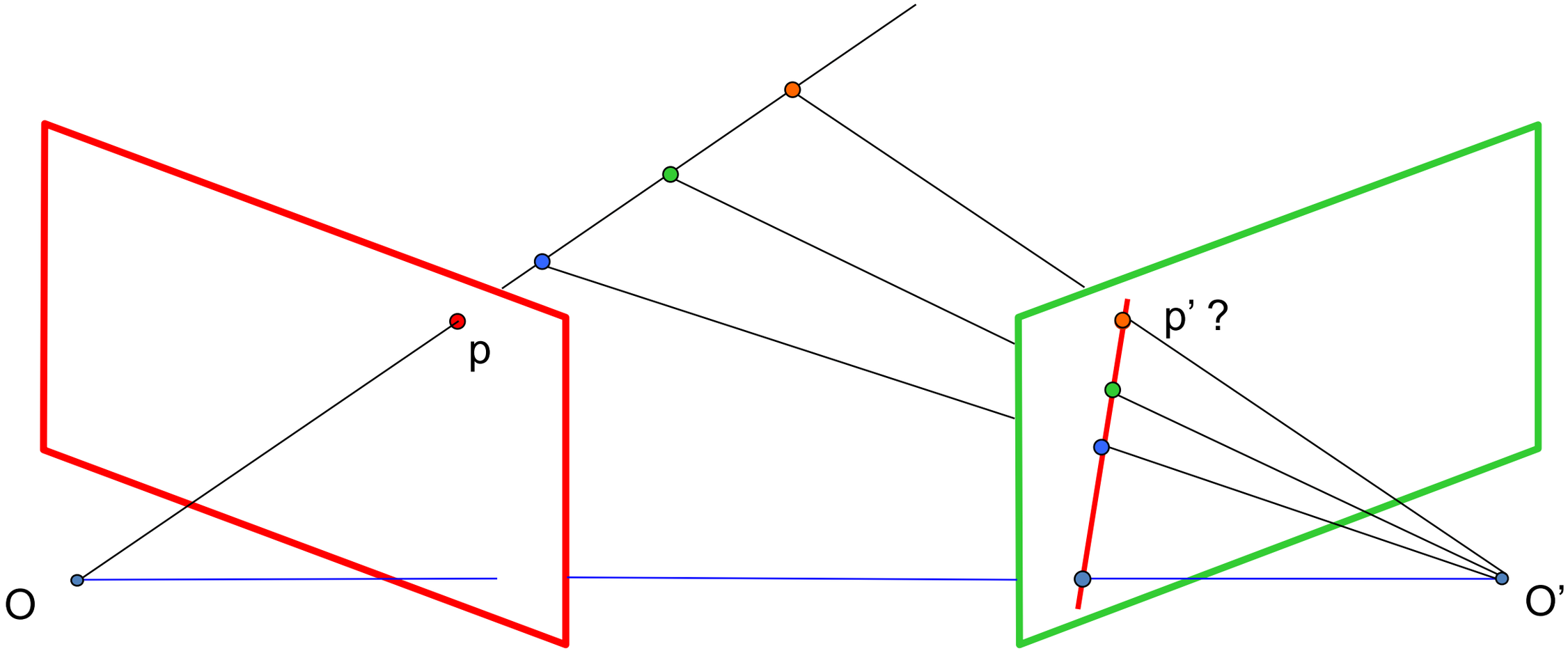
Stereo correspondence constraints



If we see a point in camera 1, are there any constraints on where we will find it on camera 2?

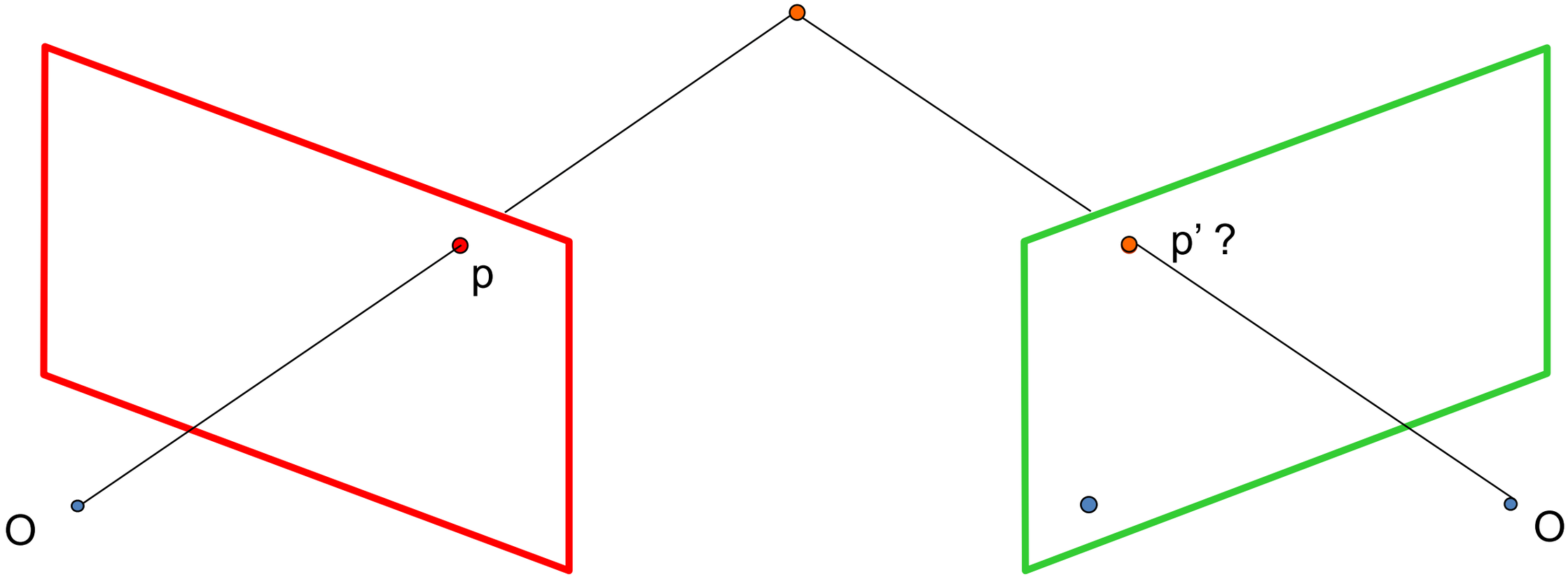
Slide credit: Antonio Torralba

Epipolar constraint



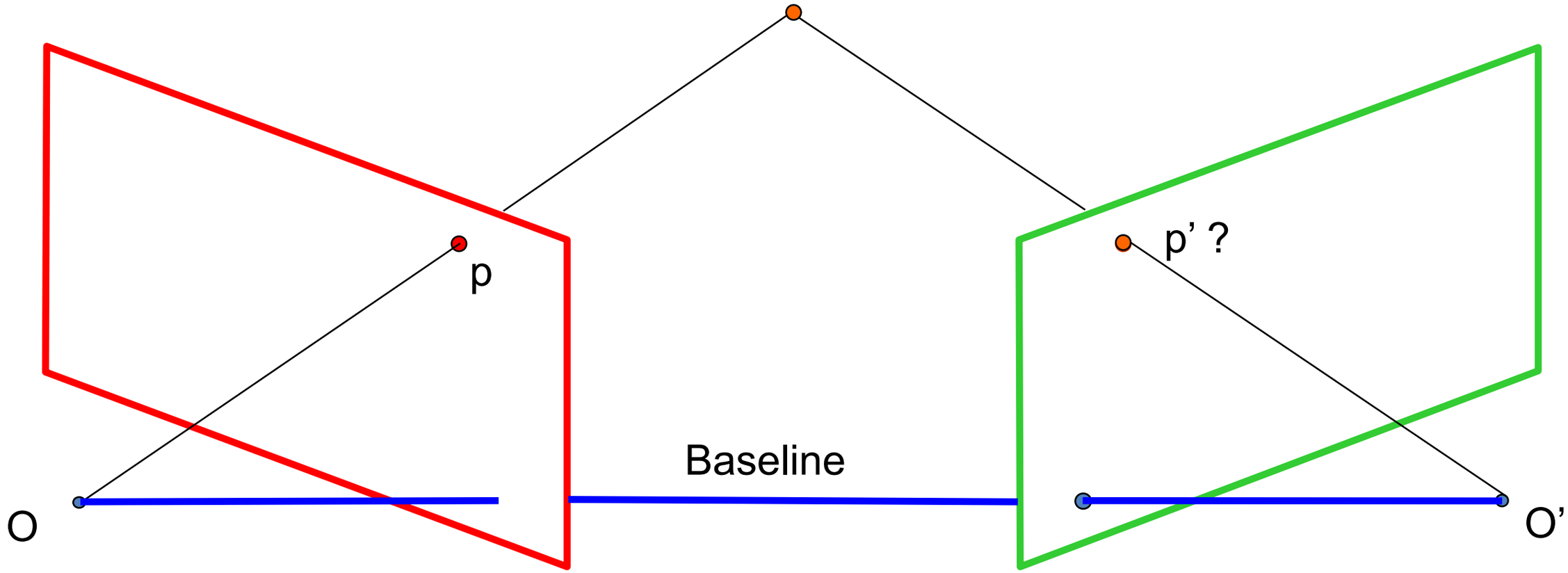
Slide credit: Antonio Torralba

Some terminology



Slide credit: Antonio Torralba

Some terminology

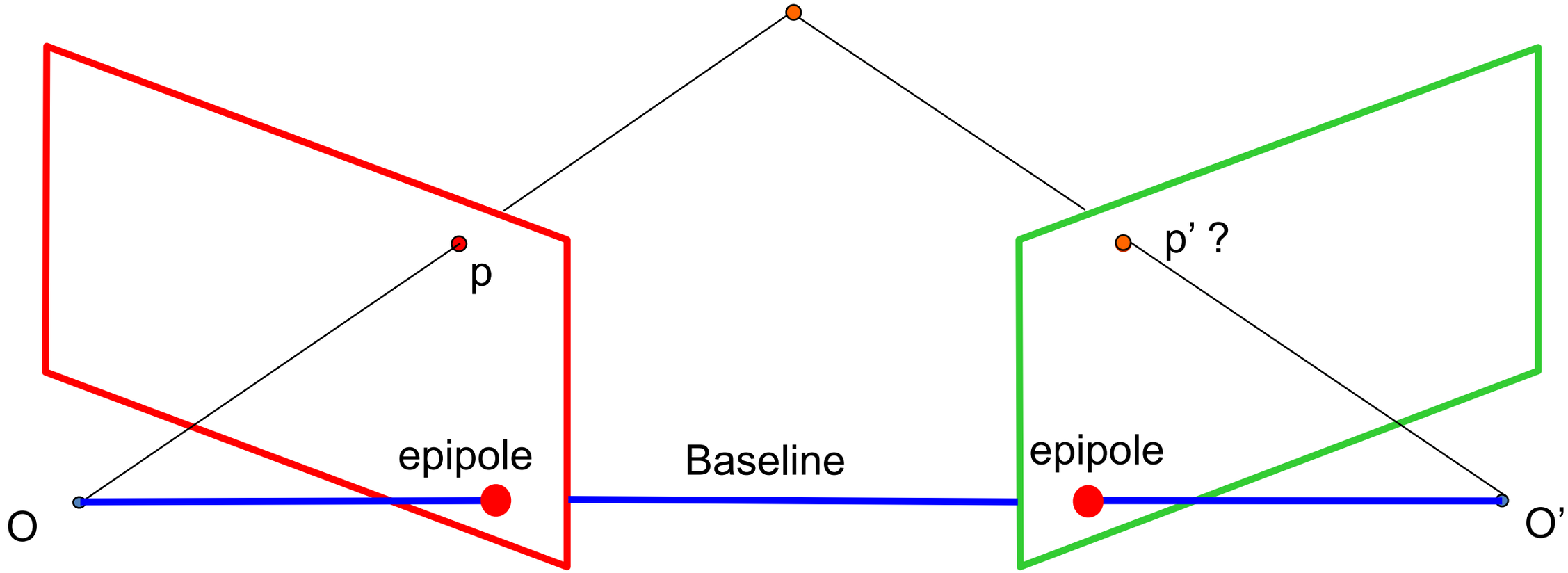


Baseline: the line connecting the two camera centers

Epipole: point of intersection of *baseline* with the image plane

Slide credit: Antonio Torralba

Some terminology

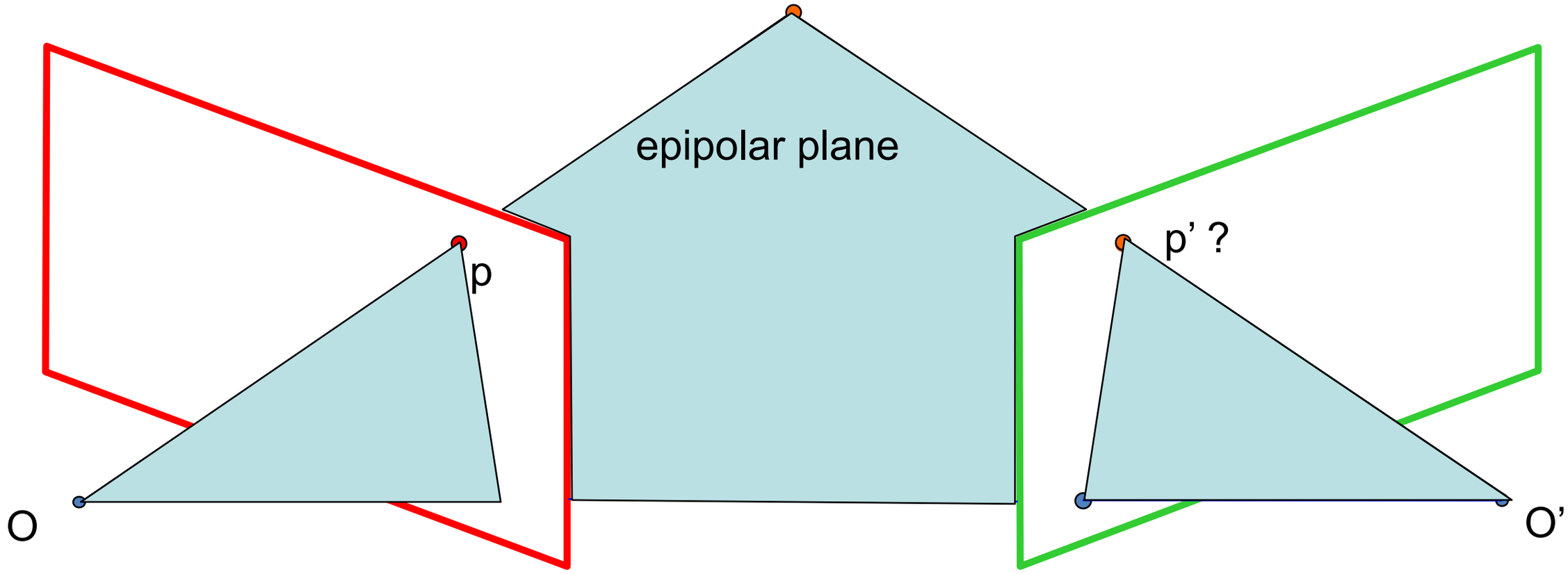


Baseline: the line connecting the two camera centers

Epipole: point of intersection of *baseline* with the image plane

Slide credit: Antonio Torralba

Some terminology



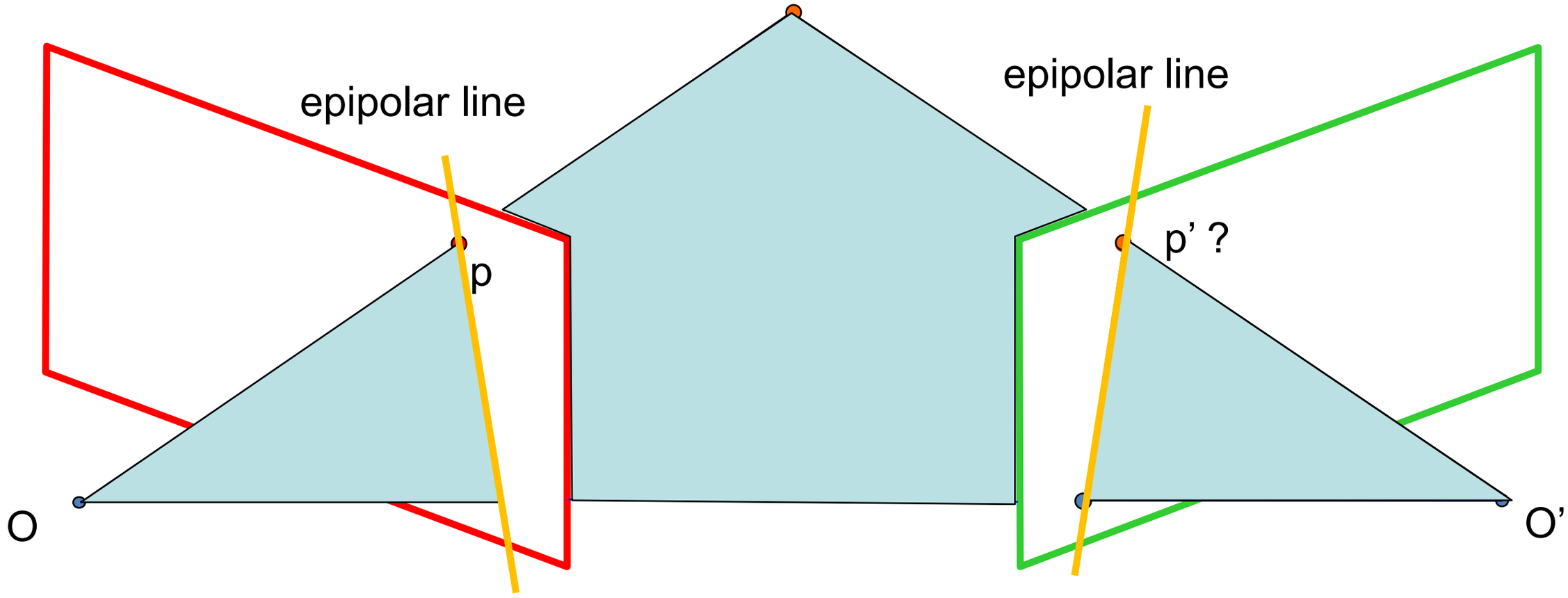
Baseline: the line connecting the two camera centers

Epipole: point of intersection of *baseline* with the image plane

Epipolar plane: the plane that contains the two camera centers and a 3D point in the world

Slide credit: Antonio Torralba

Some terminology



Baseline: the line connecting the two camera centers

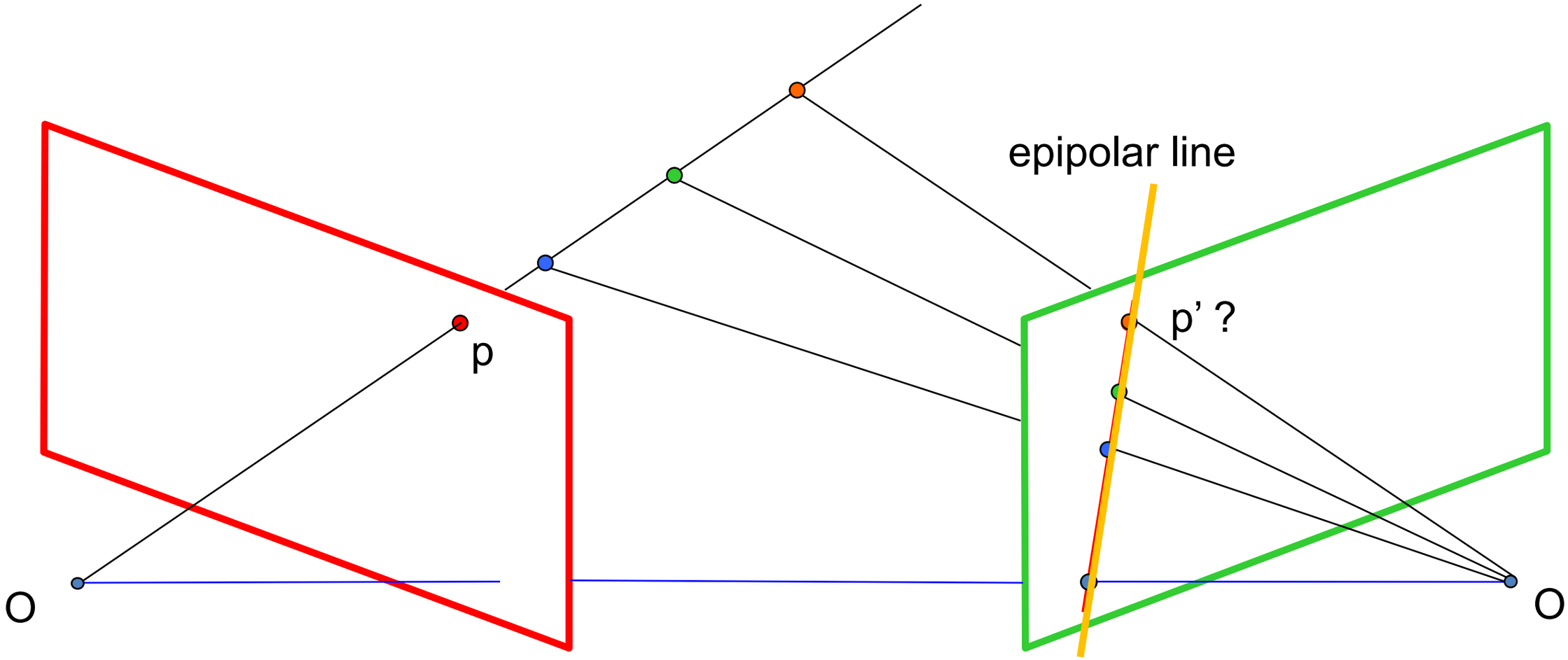
Epipole: point of intersection of *baseline* with the image plane

Epipolar plane: the plane that contains the two camera centers and a 3D point in the world

Epipolar line: intersection of the *epipolar plane* with each image plane

Slide credit: Antonio Torralba

Epipolar constraint

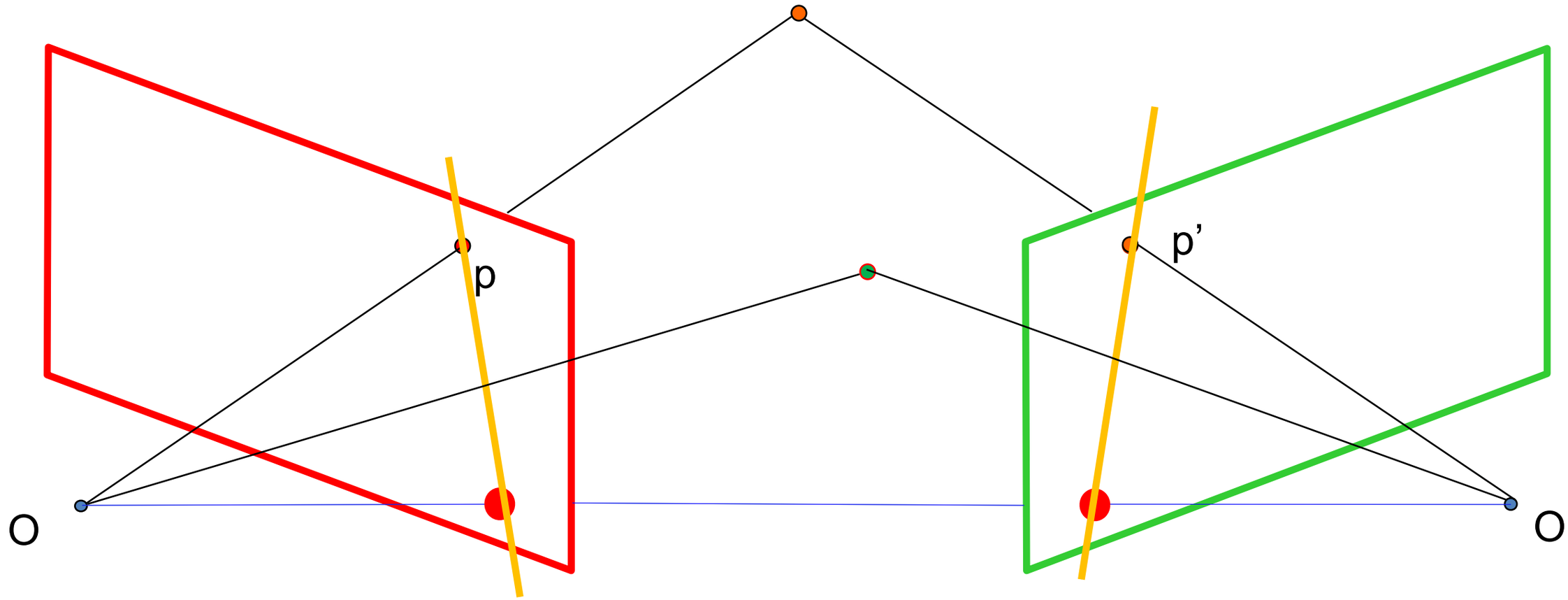


We can search for matches across epipolar lines

All epipolar lines intersect at the epipoles

Slide credit: Antonio Torralba

The essential matrix



If we observe a point in one image, its position in the other image is constrained to lie on line defined by above.

$$p^T E p' = 0$$

E: essential matrix
p, p’: image points in homogeneous coordinates

Slide credit: Antonio Torralba

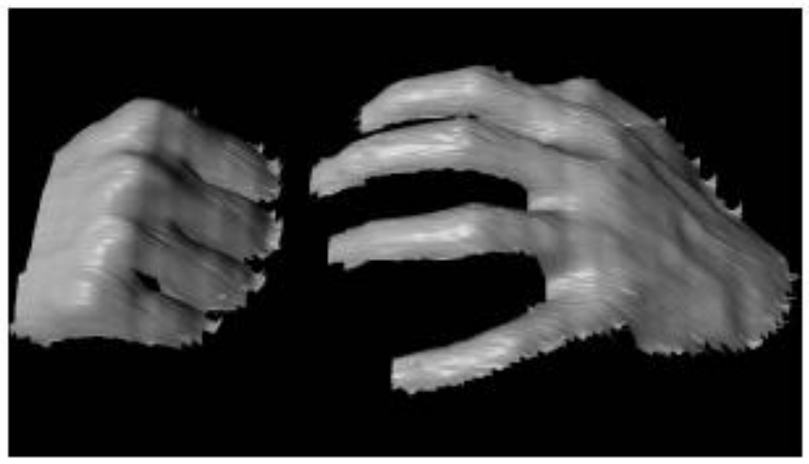
Real-time stereo

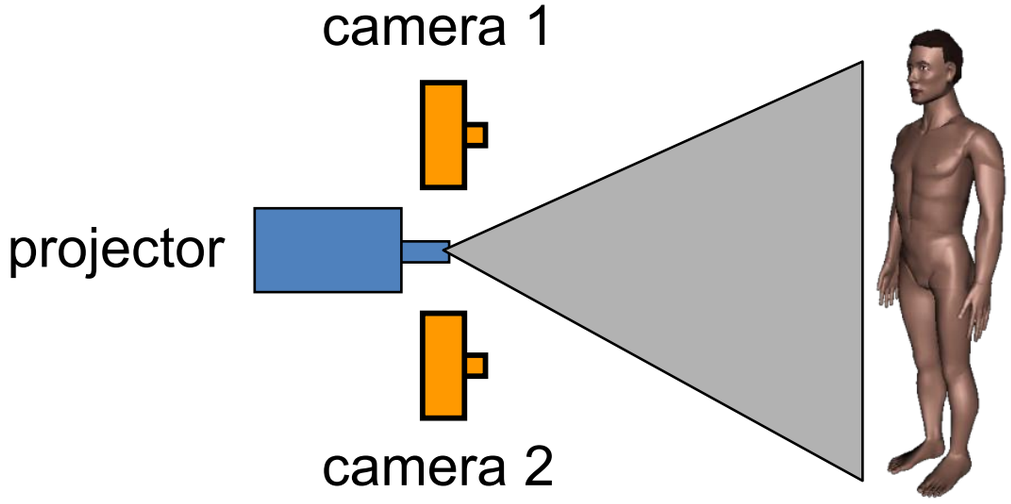
Nomad robot searches for meteorites in Antartica
http://www.frc.ri.cmu.edu/projects/meteorobot/index.html

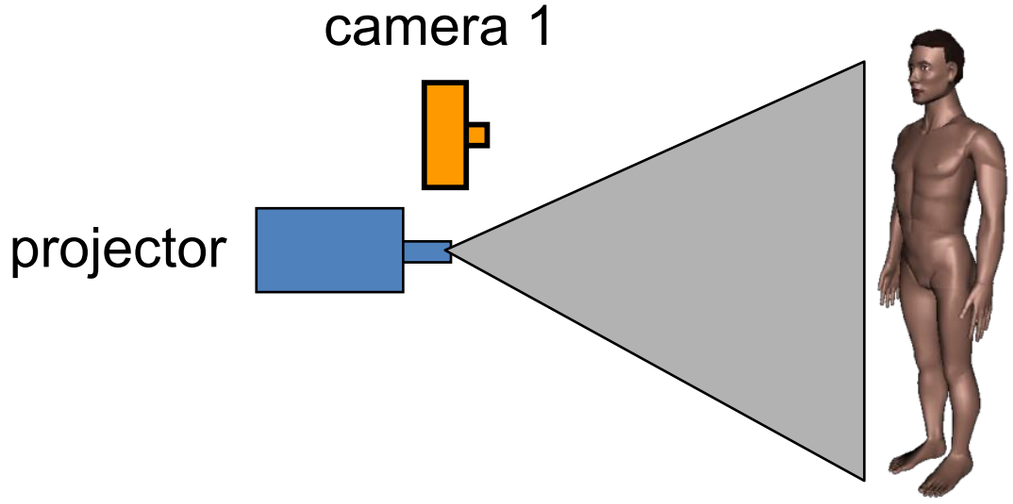
- Used for robot navigation (and other tasks)
  - Several real-time stereo techniques have been developed (most based on simple discrete search)

Active stereo with structured light

Li Zhang's one-shot stereo





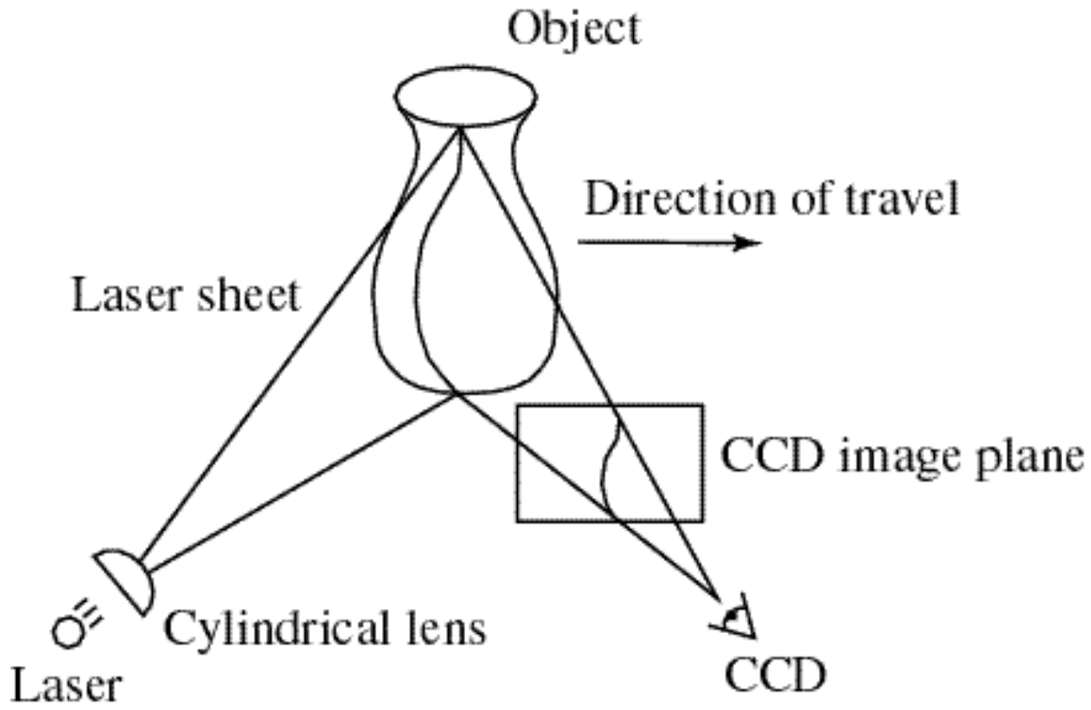


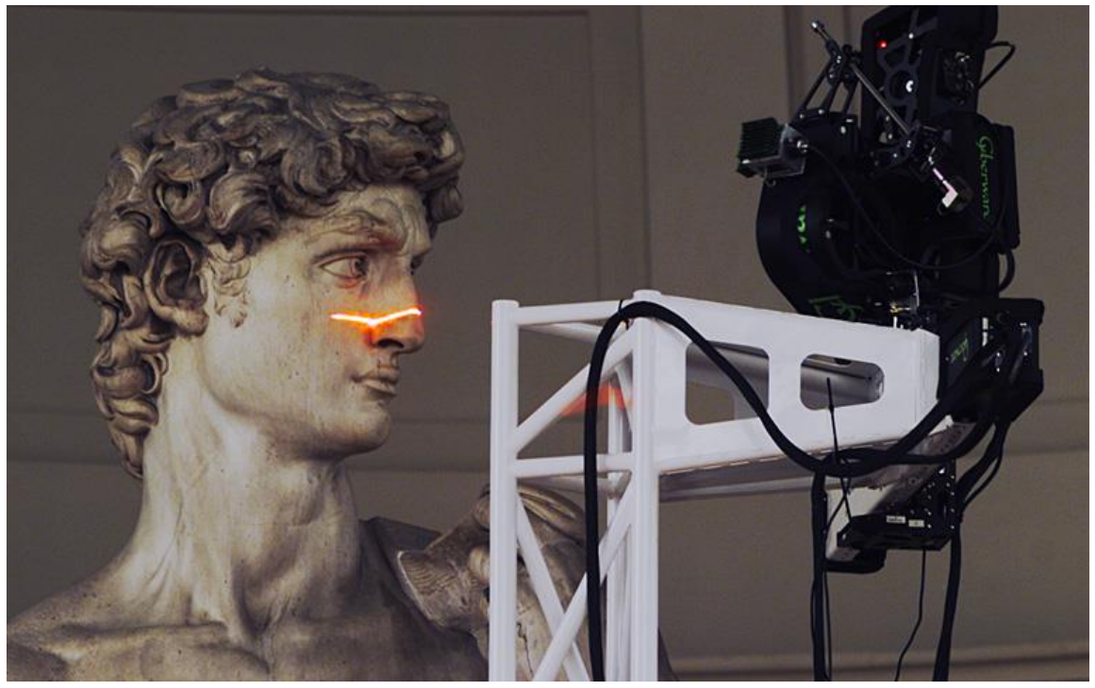
- Project "structured" light patterns onto the object
  - simplifies the correspondence problem
  - basis for active depth sensors, such as Kinect and iPhone X (using IR)

Active stereo with structured light

https://ios.gadgethacks.com/news/watch-iphone-xs-30k-ir-dots-scan-your-face-0180944/

Laser scanning





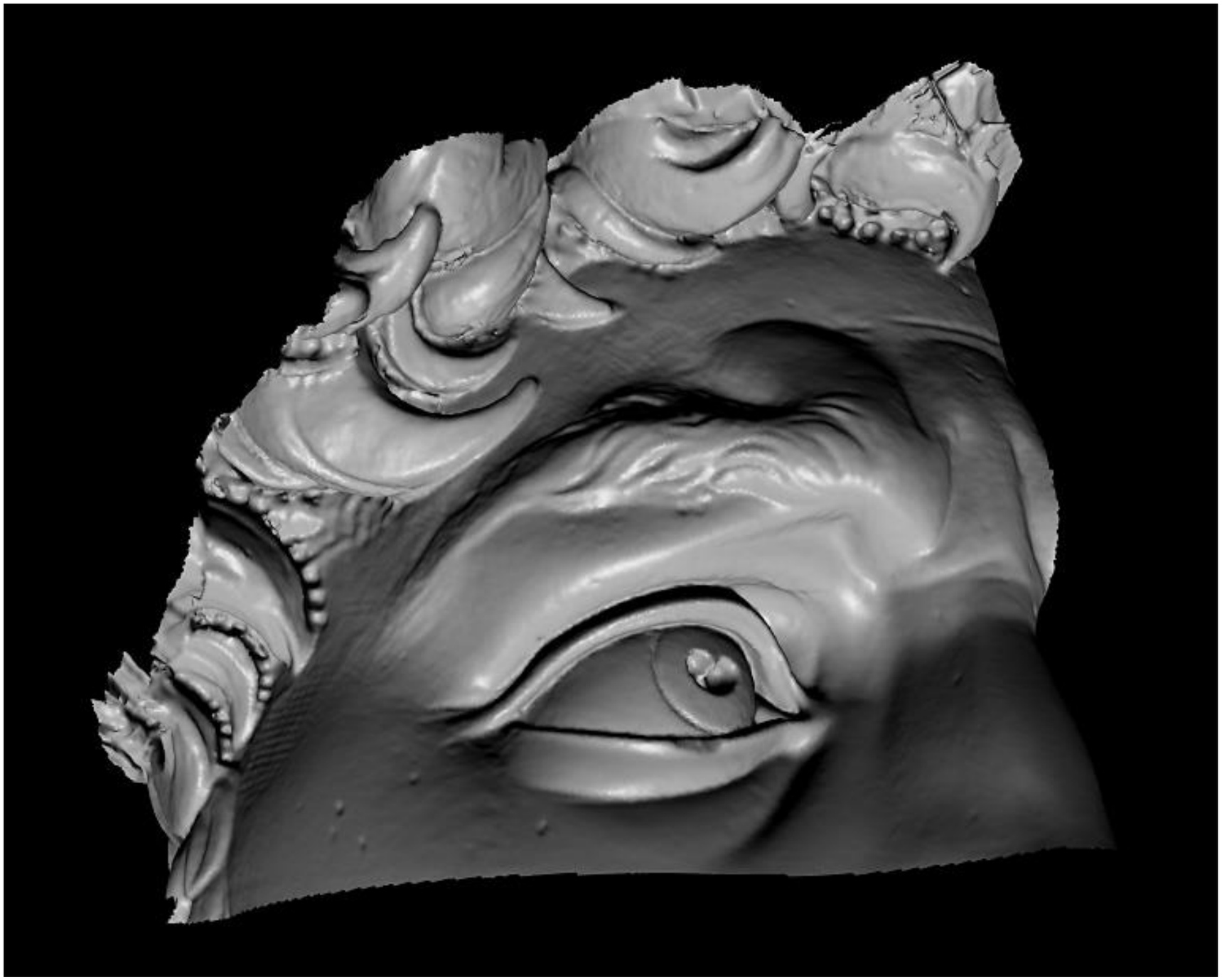
Digital Michelangelo Project
http://graphics.stanford.edu/projects/mich/

- Optical triangulation
  - Project a single stripe of laser light
  - Scan it across the surface of the object
  - This is a very precise version of structured light scanning

Laser scanned models

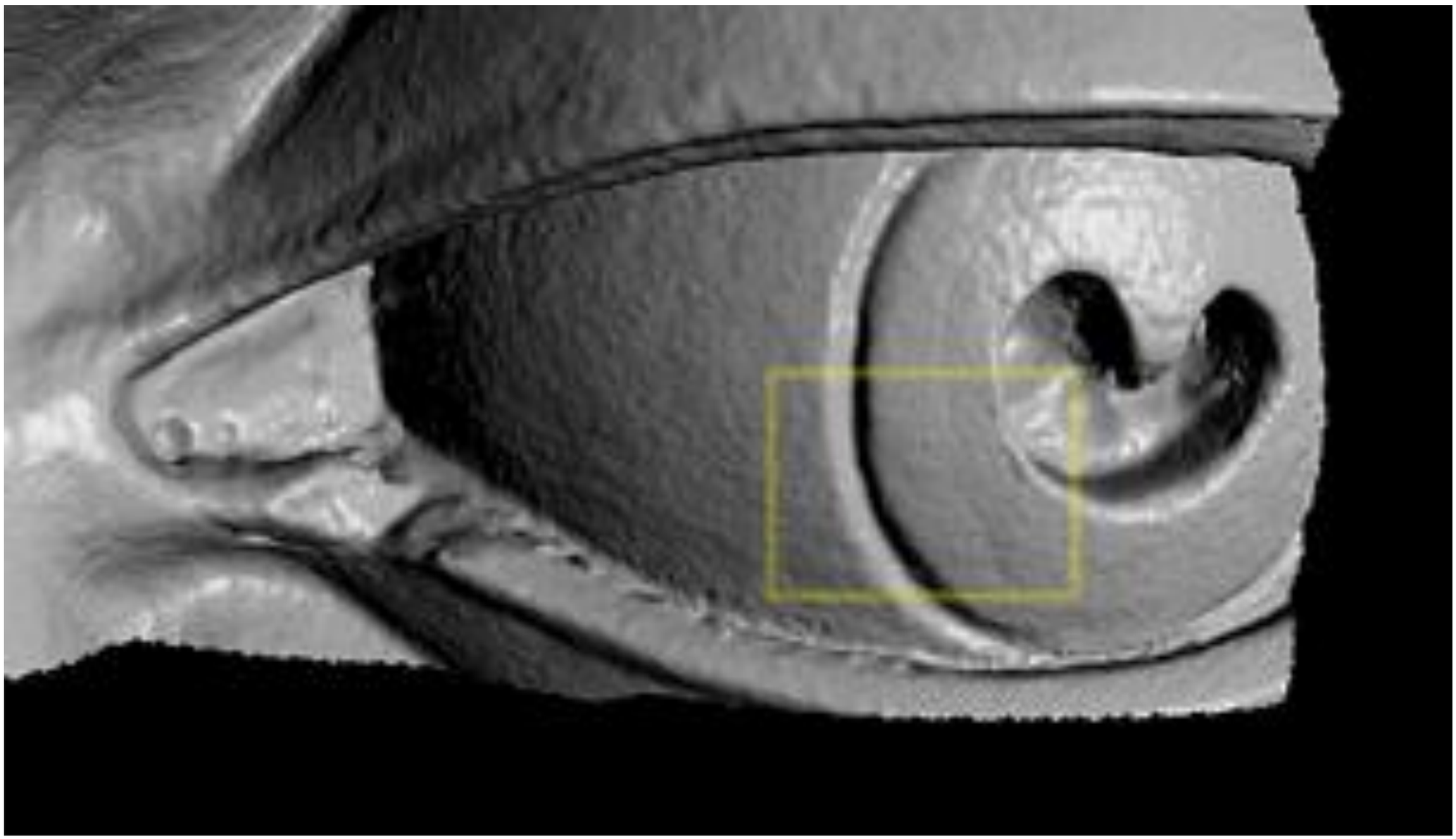
The Digital Michelangelo Project, Levoy et al.

Laser scanned models



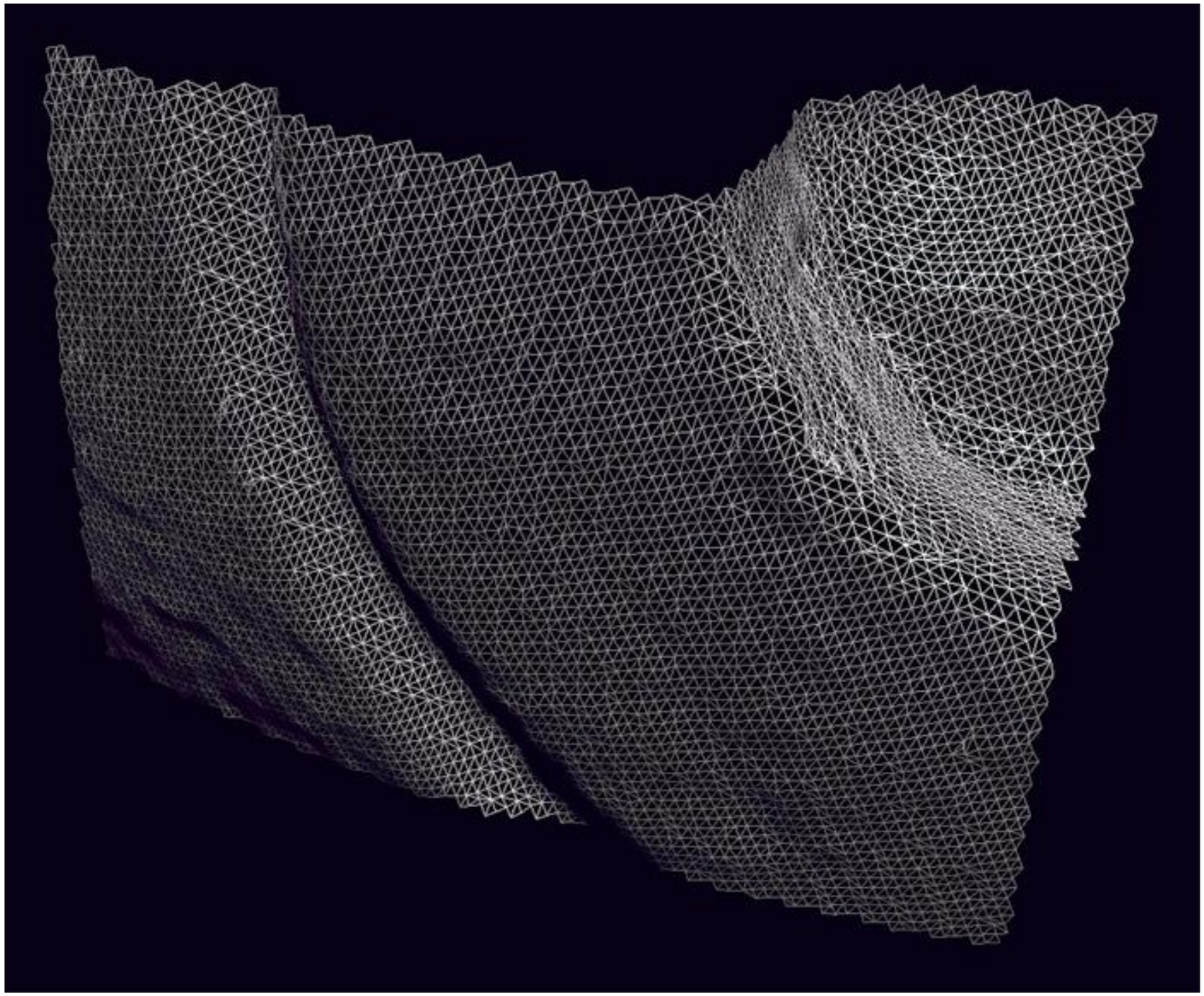
The Digital Michelangelo Project, Levoy et al.

Laser scanned models



The Digital Michelangelo Project, Levoy et al.

Laser scanned models



The Digital Michelangelo Project, Levoy et al.