SAIR

Spatial AI & Robotics Lab

CSE 473/573-A

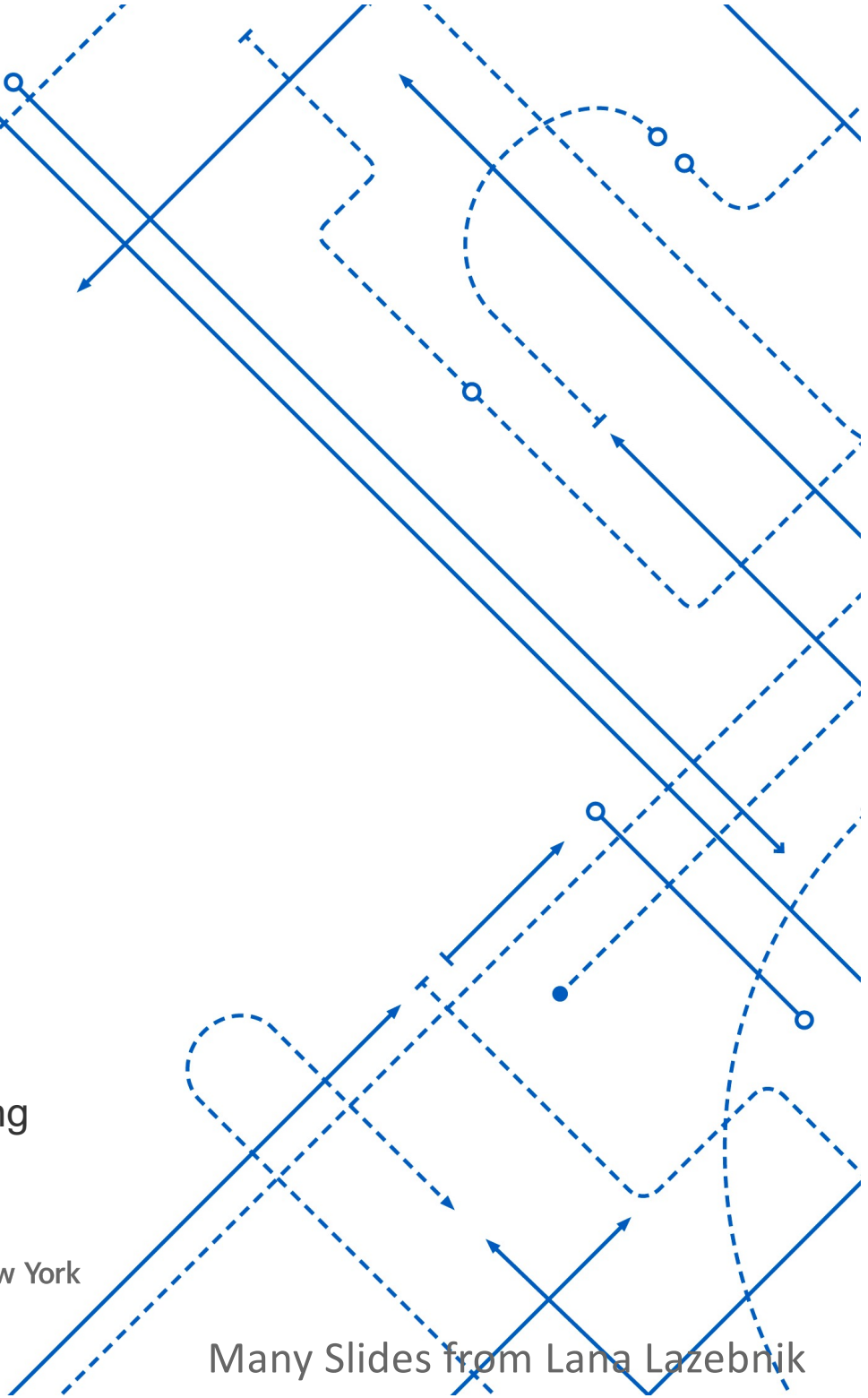
W12: STEREO VISION

Chen Wang

Spatial AI & Robotics Lab

Department of Computer Science and Engineering

University at Buffalo The State University of New York



Many Slides from Lana Lazebnik

SAIR

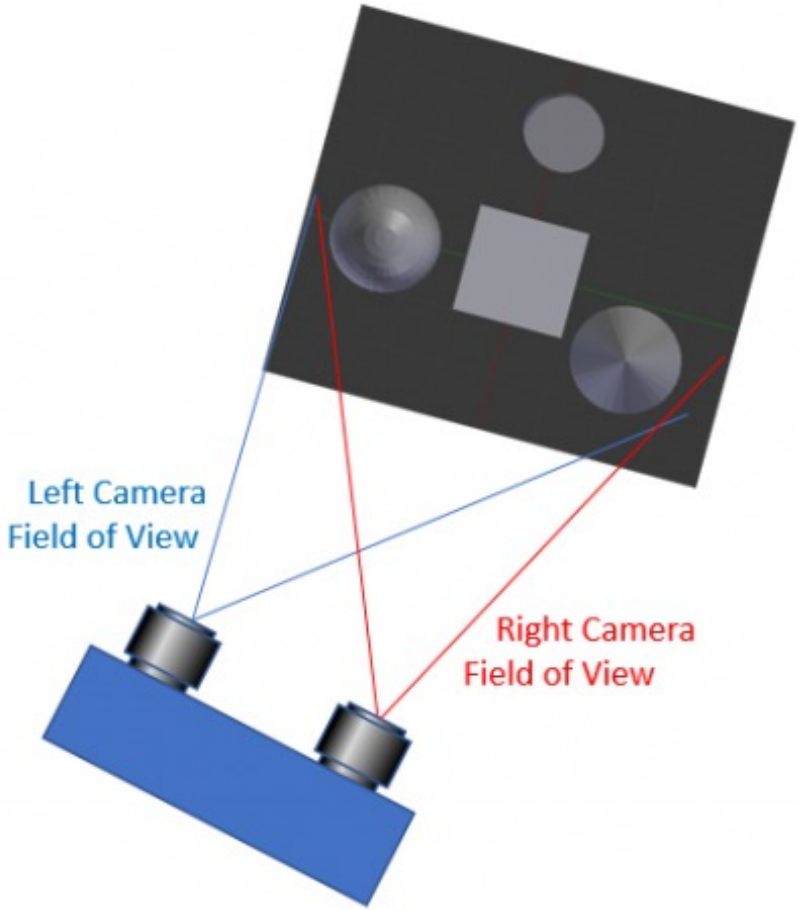
Spatial AI & Robotics Lab

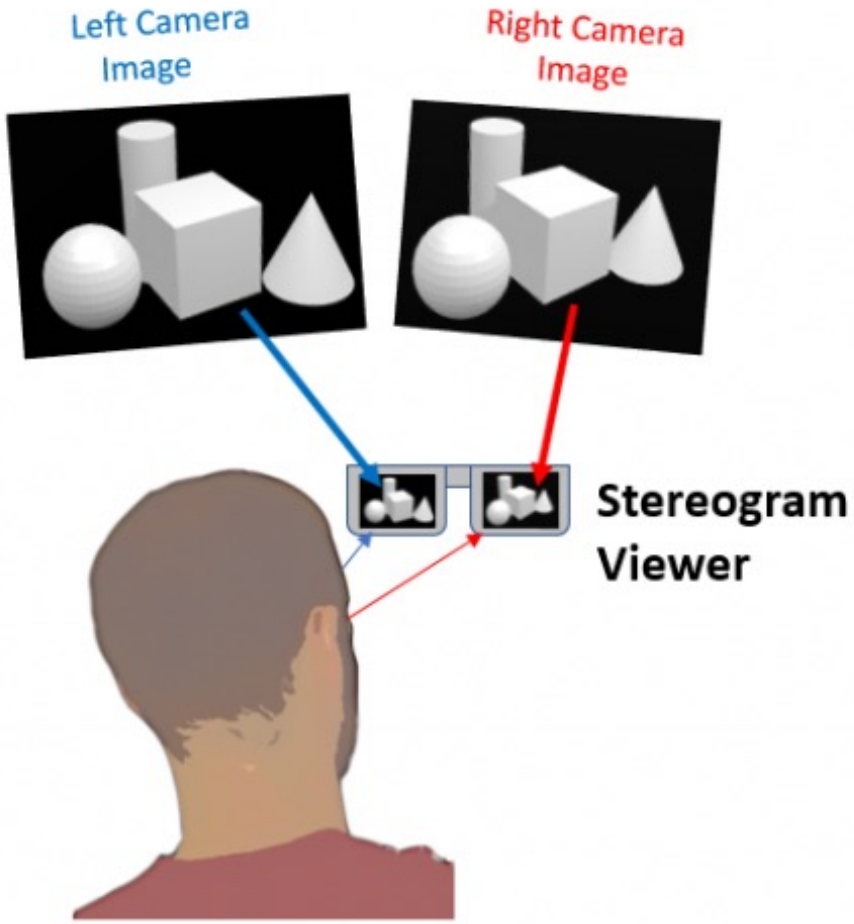
STEREO VISION

Epipolar geometry

Stereo Vision (binocular camera)

Stereo Vision



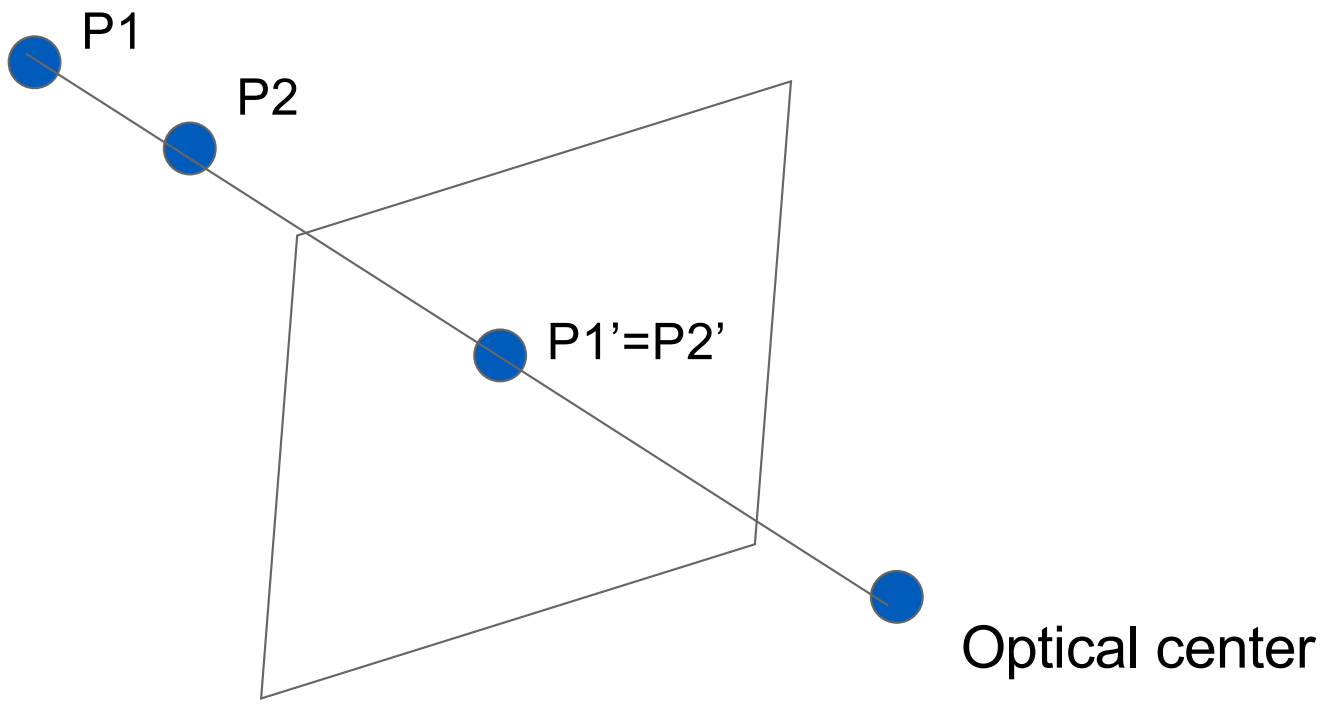


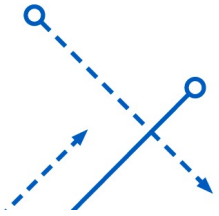
Why multiple views?

- Structure and depth are inherently ambiguous from single views.

Why multiple views?

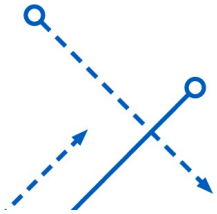
- Structure and depth are inherently ambiguous from single views.





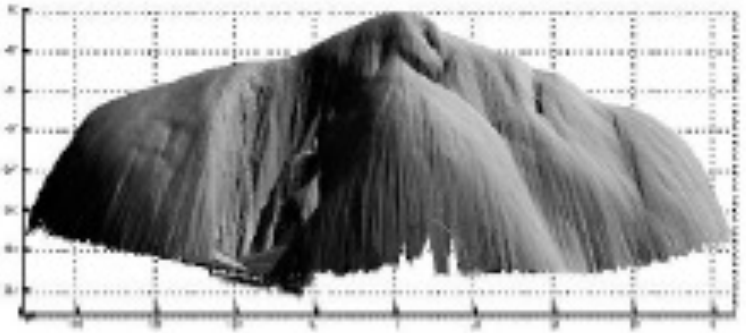
Stereo Vision

- What cues help perceive 3D shape and depth?
  - Shading
  - Focus/Defocus
  - Texture
  - Perspective
  - Motion
  - Occlusion

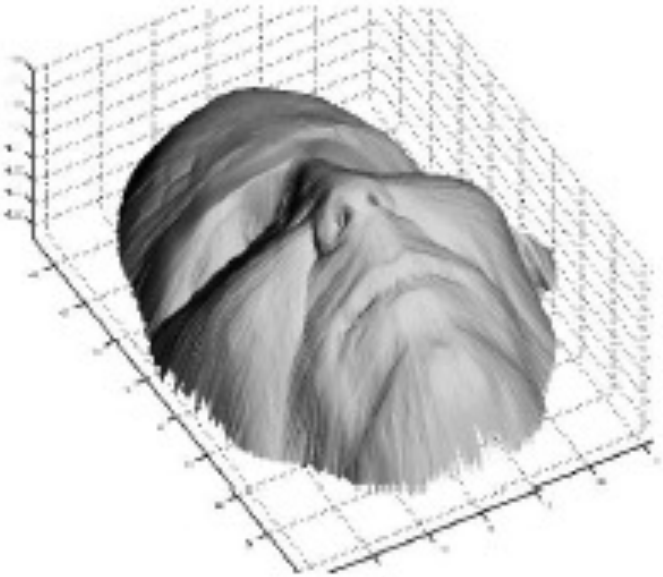


Shading

a)



b)

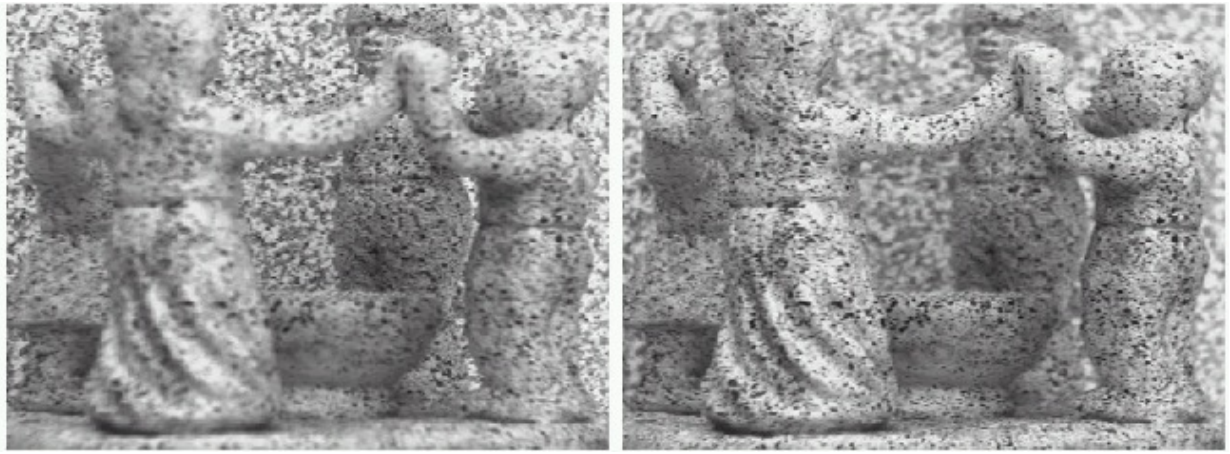


c)

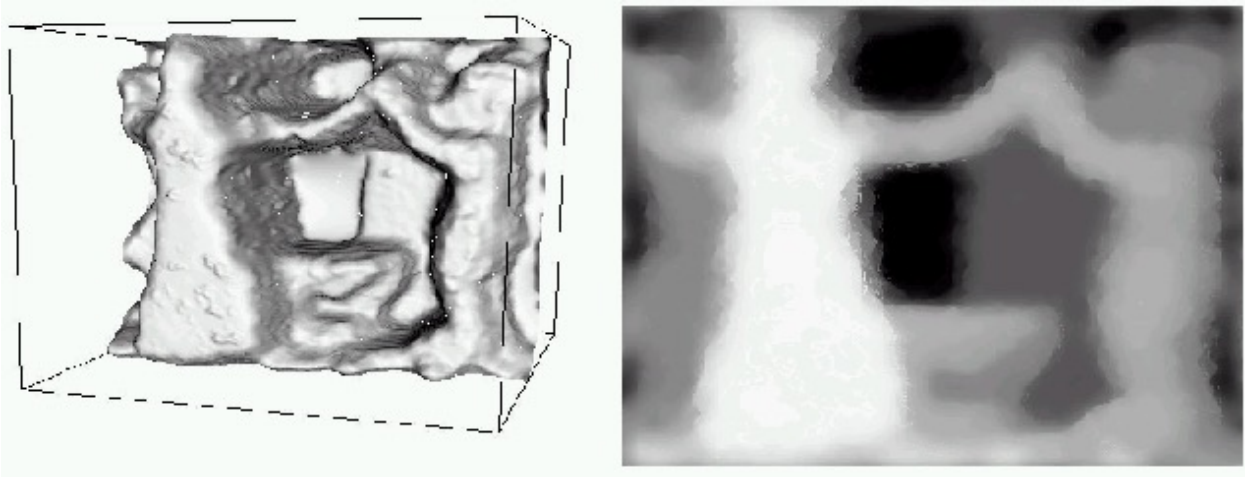
Prados & Faugeras 2006

Focus/defocus

- Same point of view, different camera parameters

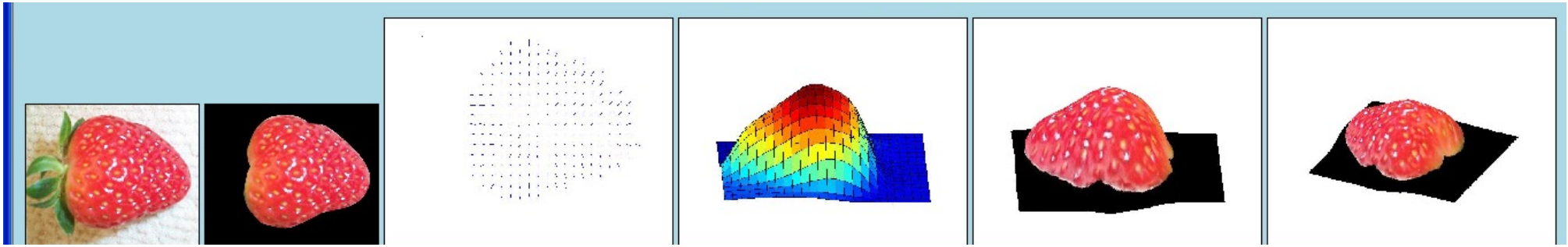


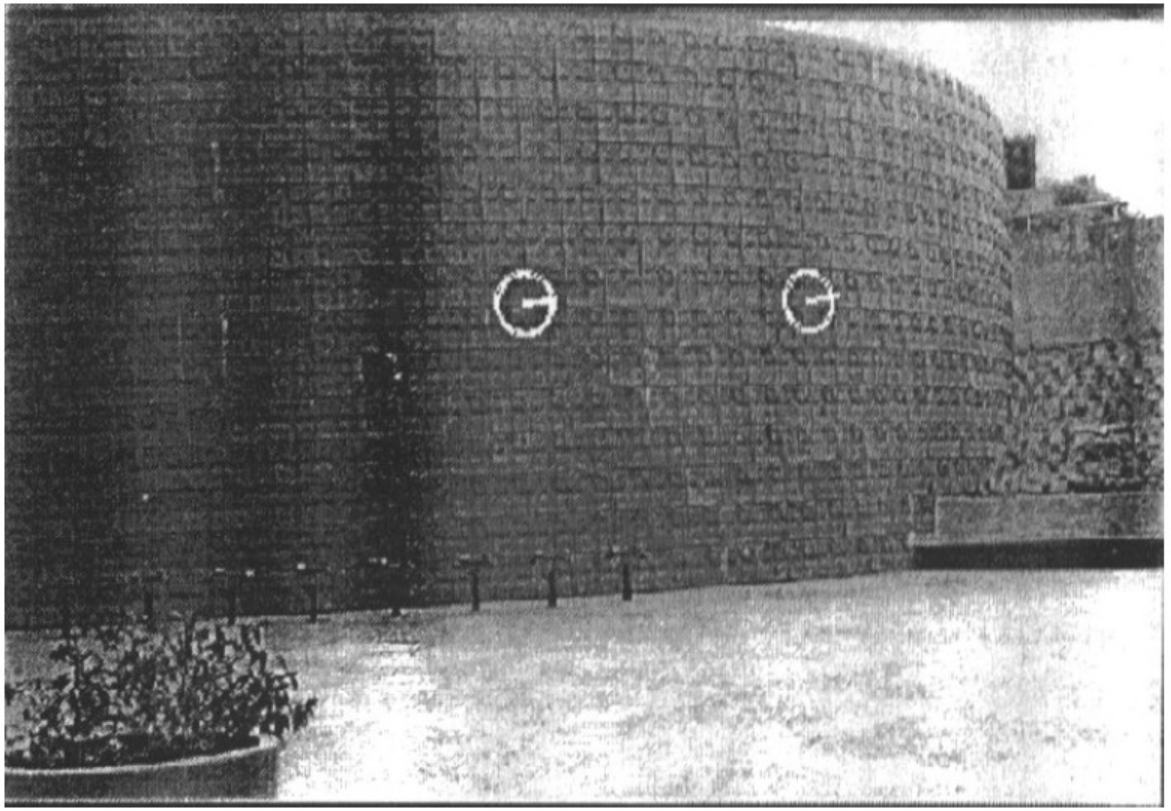
- 3D shape / depth estimates



H. Jin and P. Favaro, 2002

Texture





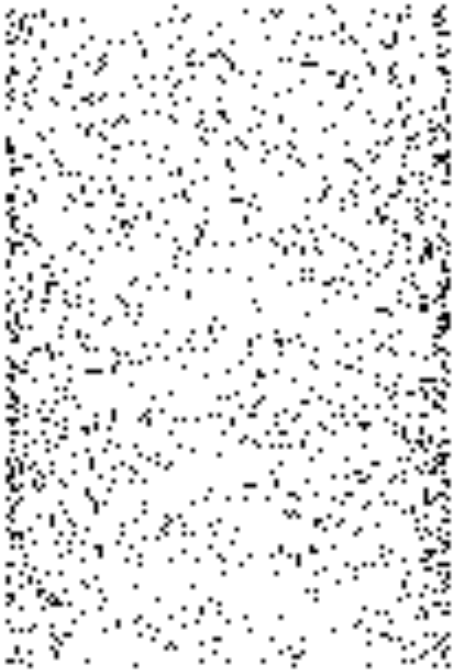
[From A.M. Loh. The recovery of 3-D structure using visual texture patterns. PhD thesis]

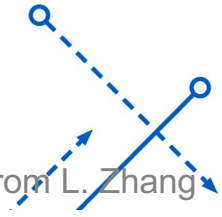
Perspective effects

Image credit: S. Seitz

Motion



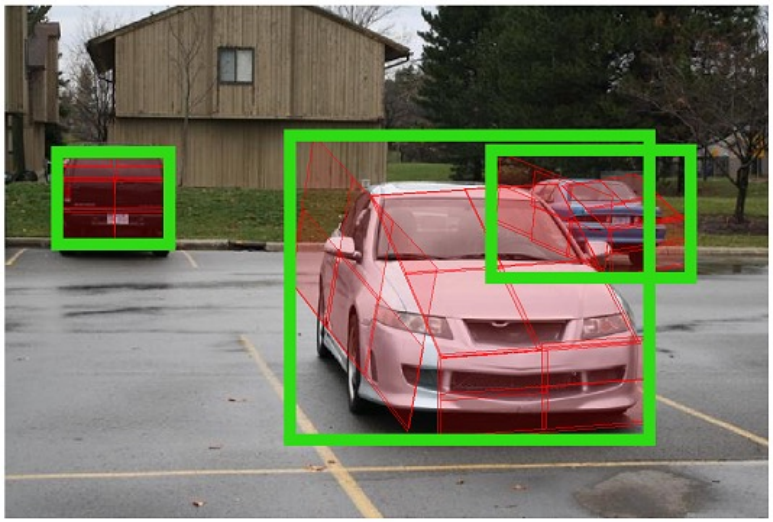





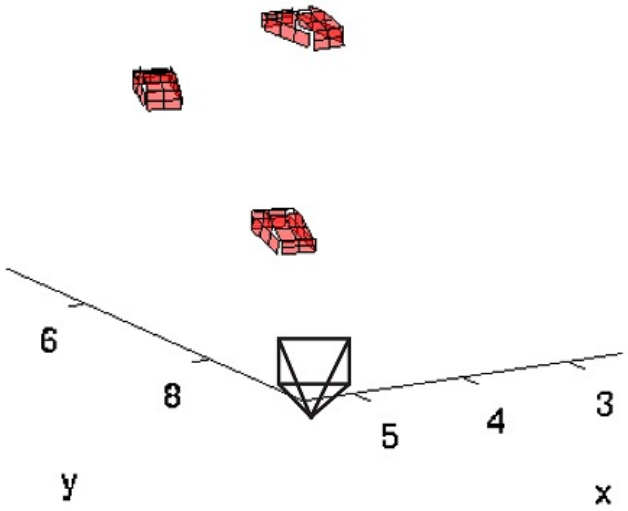
Figures from L. Zhang

Occlusion

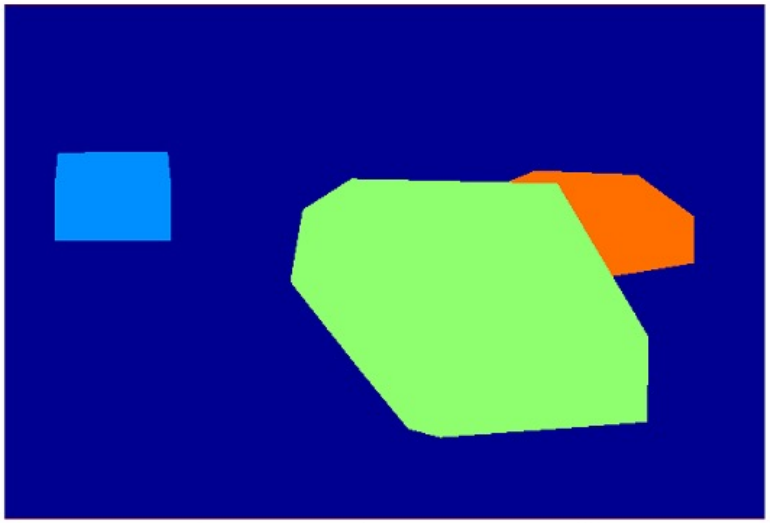
(a) input image



(b) 2D detection



(c) 3D spatial layout



(d) 2D object mask

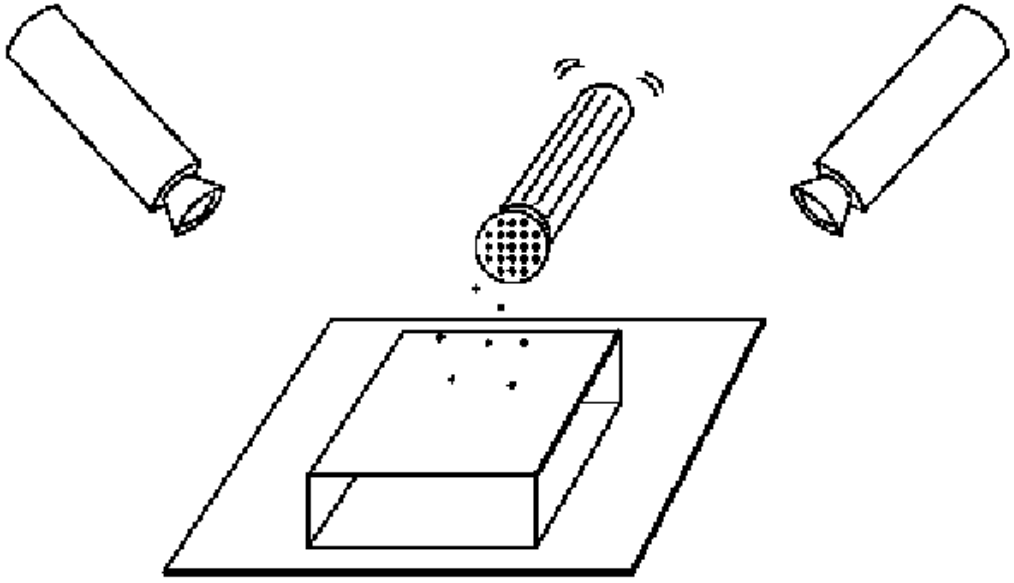
Yu Xiang and Silvio Savarese. Object Detection by 3D Aspectlets and Occlusion Reasoning. In the 4th International IEEE Workshop on 3D Representation and Recognition (3dRR), 2013

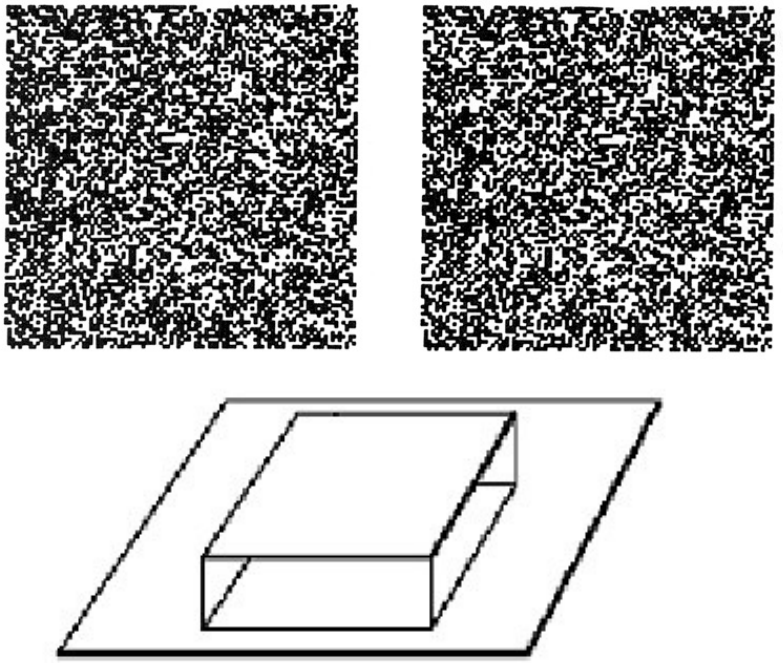
Animal Binocular Systems

- If stereo is critical for depth perception, navigation, recognition, etc., then this would be a problem

Random dot stereograms

- Julesz 1960: Do we identify local brightness patterns before fusion (monocular) or after (binocular)?
- To test: pair of synthetic images obtained by randomly spraying black dots on white objects

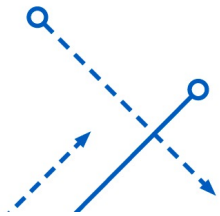




From Palmer, "Vision Science", MIT Press

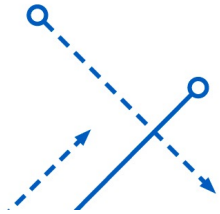
Random dot stereograms

- When viewed monocularly, they appear random; when viewed stereoscopically, see 3D structure.
- Human binocular fusion not directly associated with the physical retinas; must involve the central nervous system.
- High level scene understanding not required for Stereo
- High level scene understanding is arguably *better* than stereo

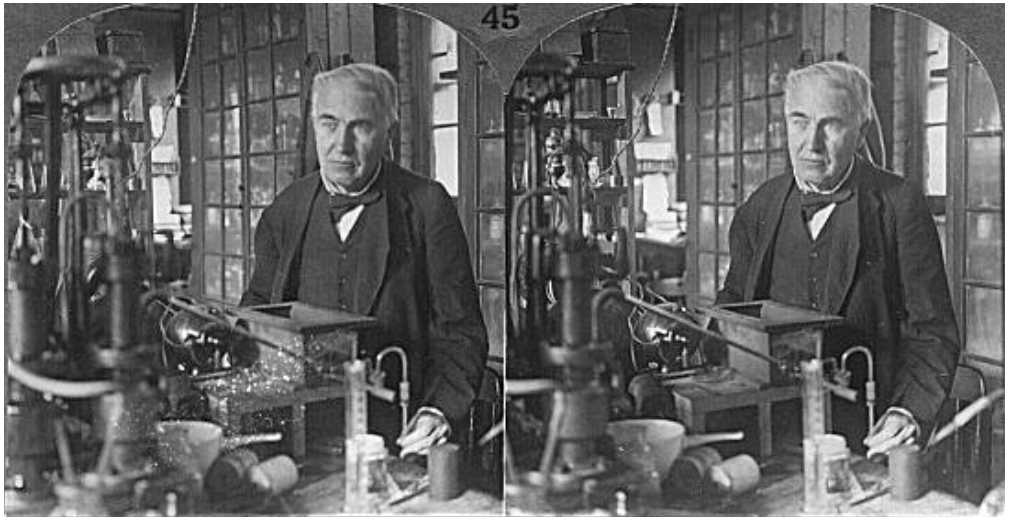


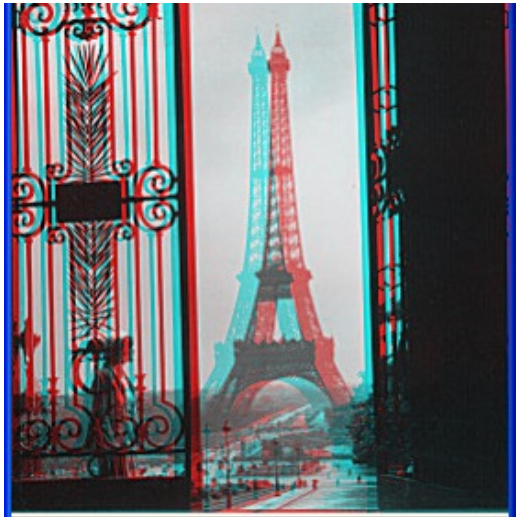
Stereo photography and stereo viewers

- Invented by Sir Charles Wheatstone, 1838
- Take two pictures of the same subject from two slightly different viewpoints and display so that each eye sees only one of the images

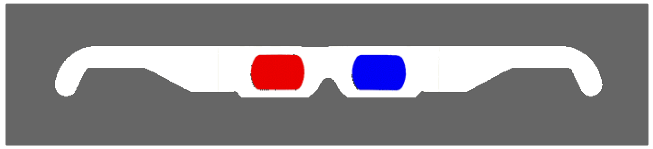


Stereo photography and stereo viewers





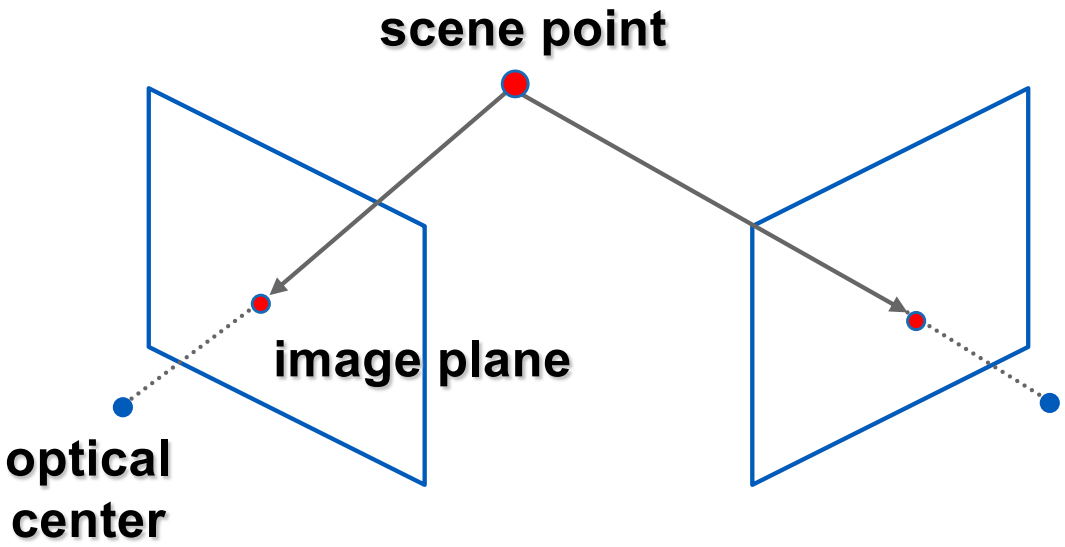
Stereo photography and stereo viewers

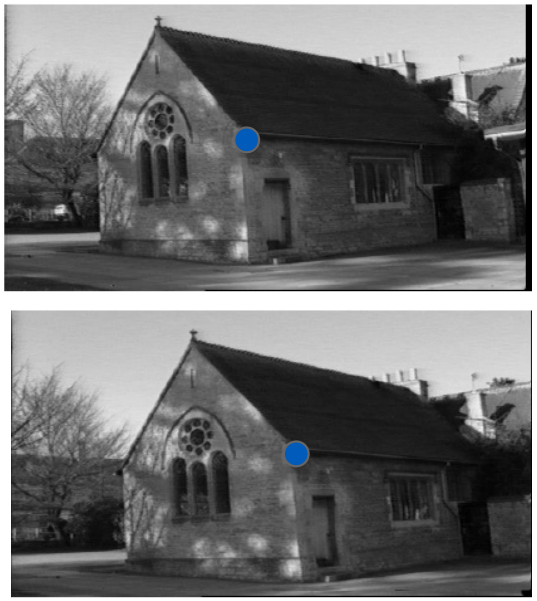


Public Library, Stereoscopic Looking Room, Chicago, by Phillips, 1923

Estimating depth with stereo

- We'll need to consider:
  - Info on camera pose ("calibration")
  - Image point correspondences



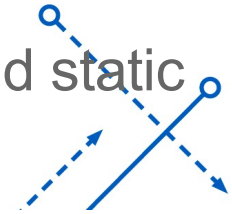


Stereo Vision

- Structure from motion
  - Shape from “motion” between two views
  - Infer 3D shape of scene from two (multiple) images from different viewpoints

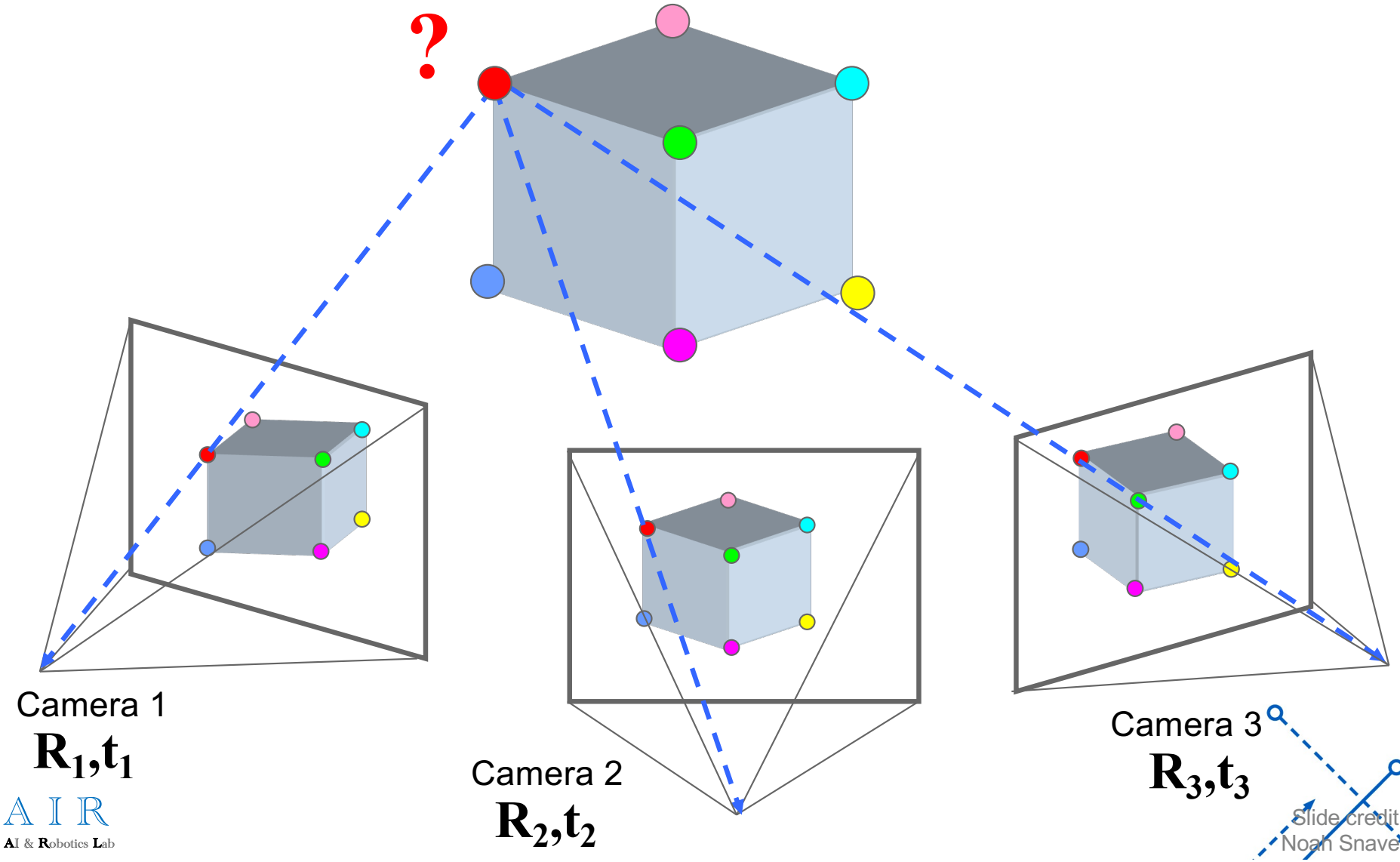
Two cameras, simultaneous views

Single moving camera and static scene



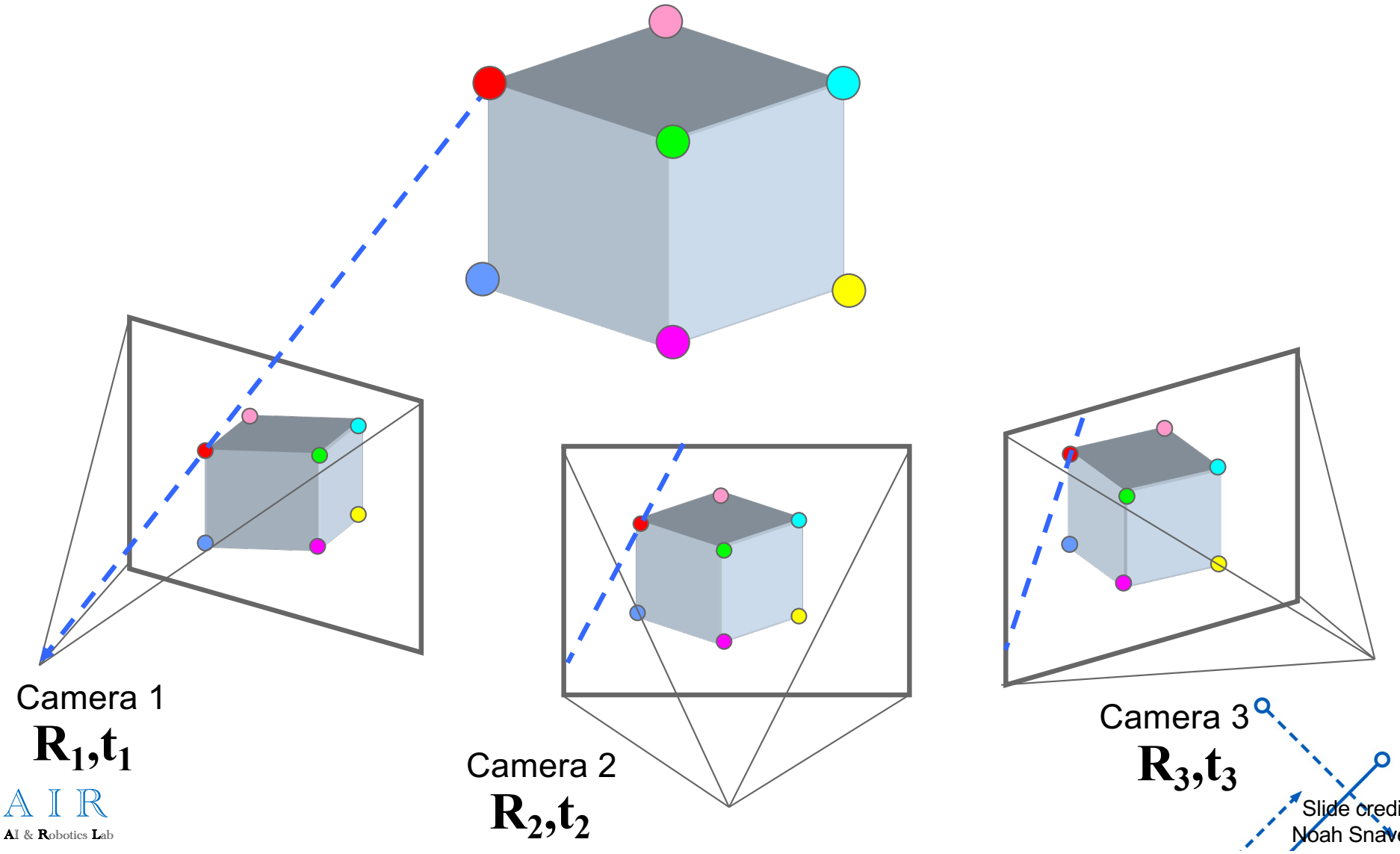
Structure from motion (SFM)

- **Structure:** Given projections of the same 3D point in two or more images, compute the 3D coordinates of that point.



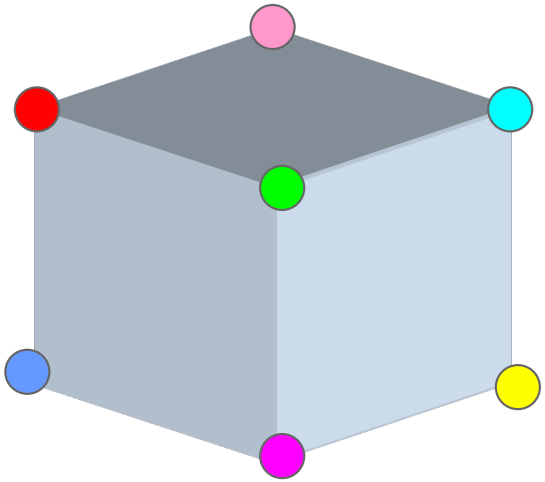
Structure from motion (SFM)

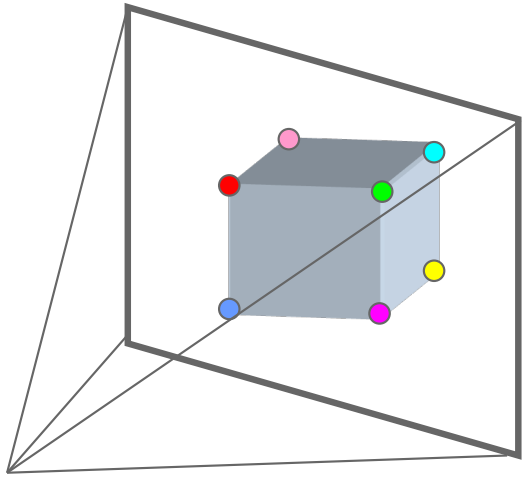
- **Stereo correspondence:** Given a point in one of the images, where could its corresponding points be in the other images?



Structure from motion (SFM)

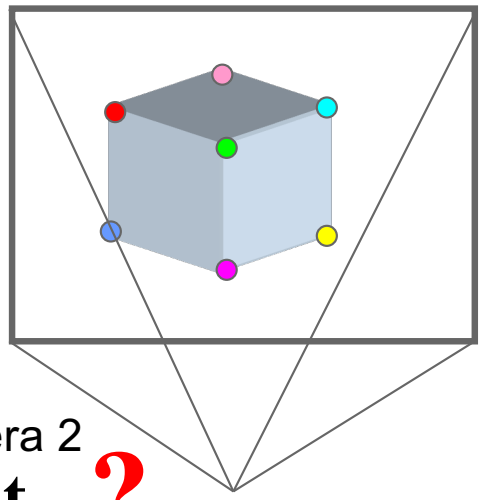
- **Motion:** Given a set of corresponding points in two or more images, compute the camera parameters





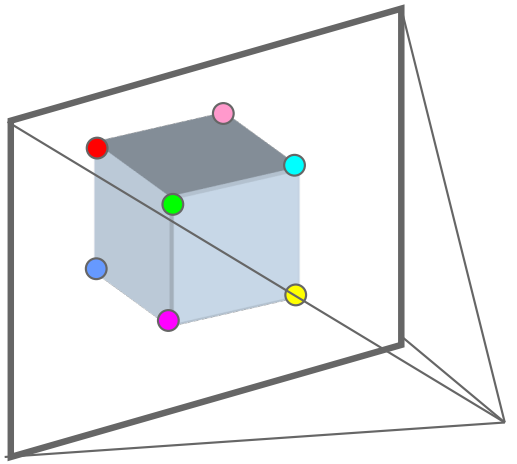
Camera 1

$\mathbf{R}_1,\mathbf{t}_1$?



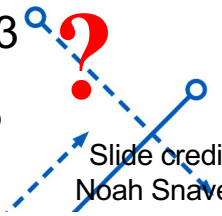
Camera 2

$\mathbf{R}_2,\mathbf{t}_2$?



Camera 3

$\mathbf{R}_3,\mathbf{t}_3$?



Slide credit: Noah Snavely

Human eye

- Rough analogy with human visual system:
  - Pupil/Iris: control light passing through lens
  - Retina: contains cells, where image is formed
  - Fovea: highest concentration of cones

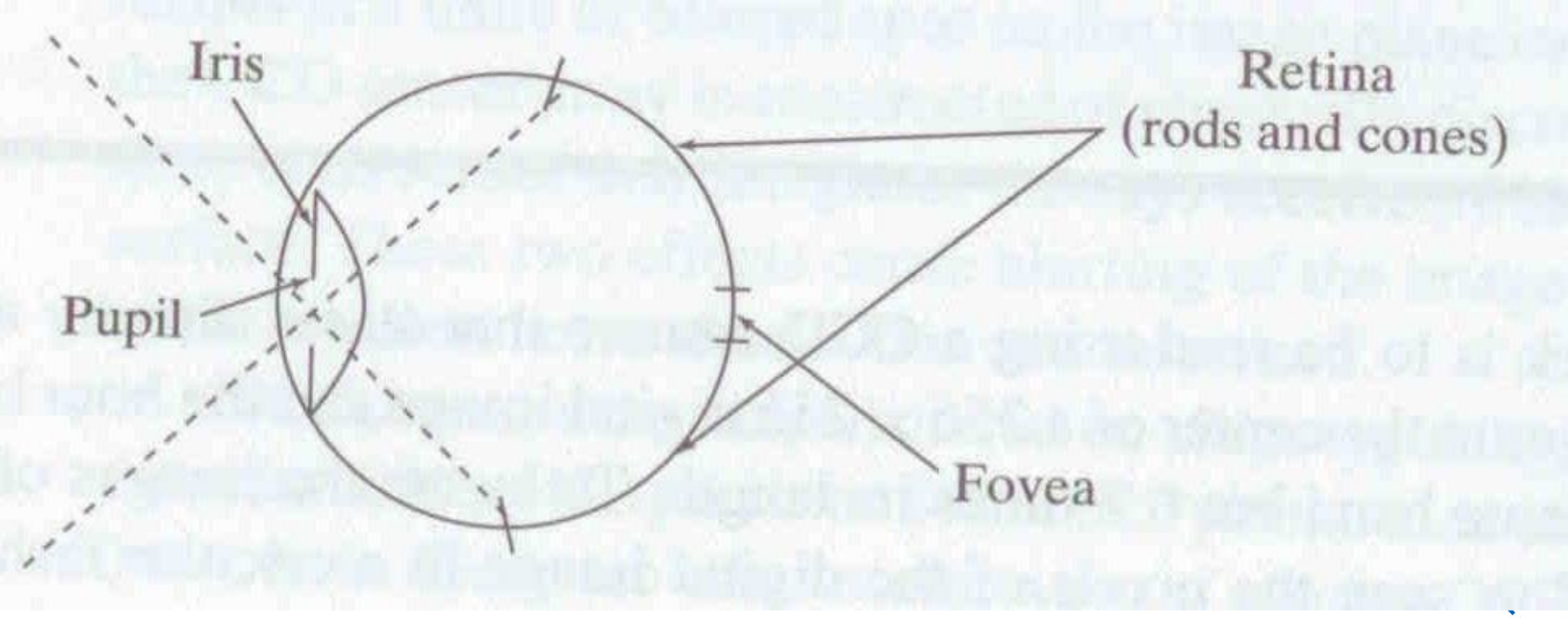
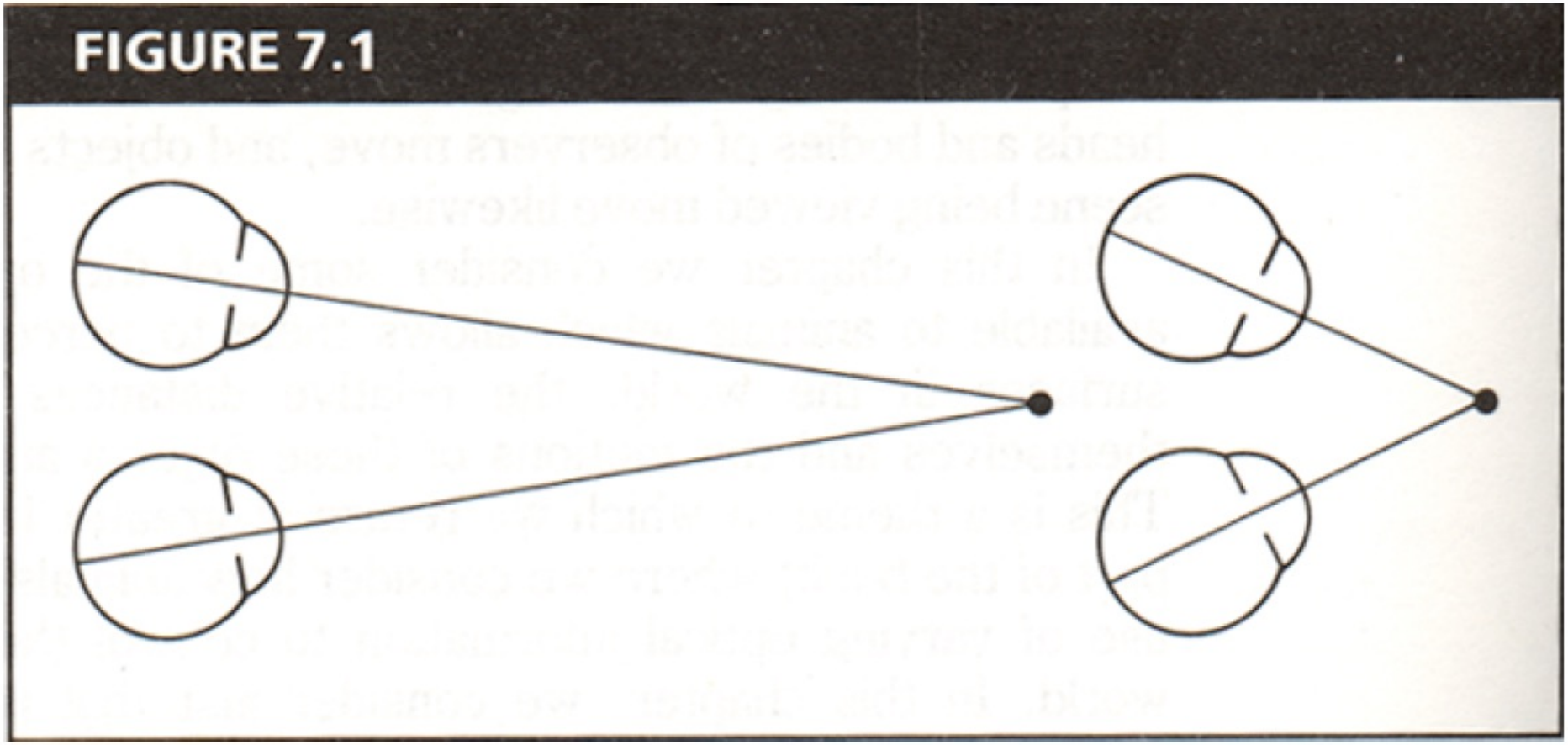


Fig from Shapiro and Stockman

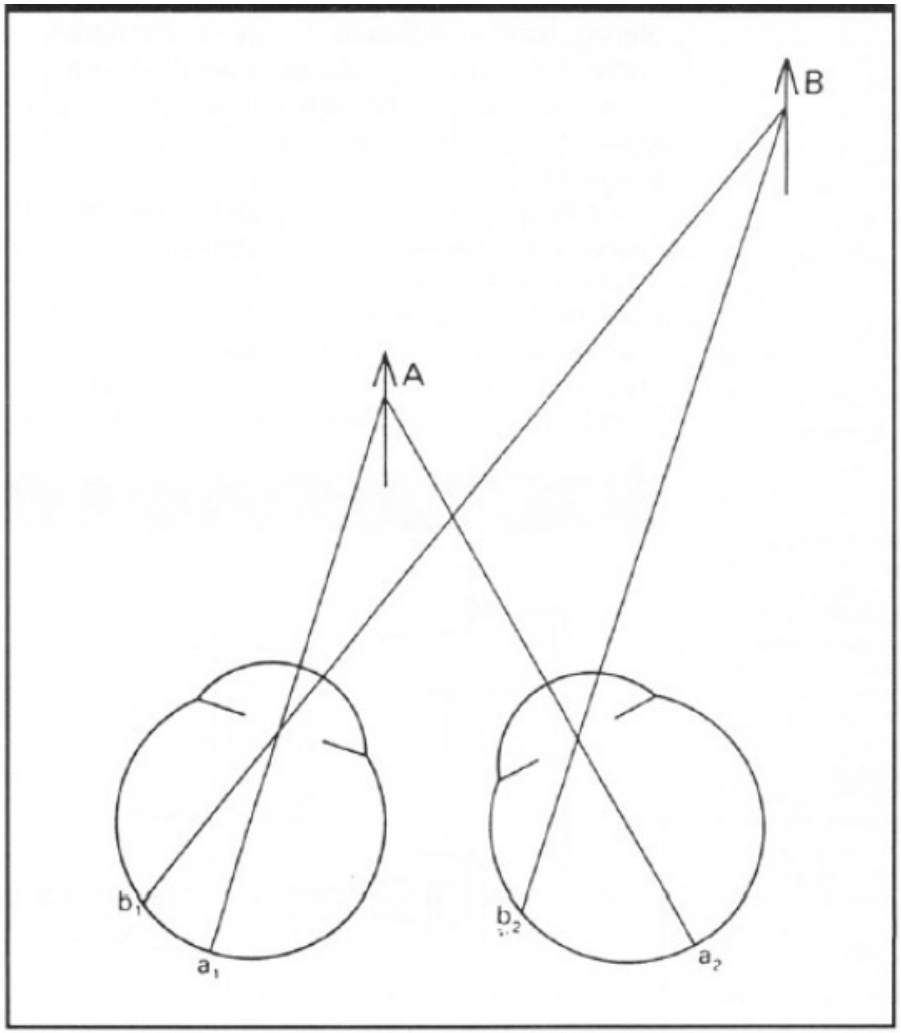
Human stereopsis: disparity

- Human eyes **fixate** on point in space – rotate so that corresponding images form in centers of fovea.



Disparity

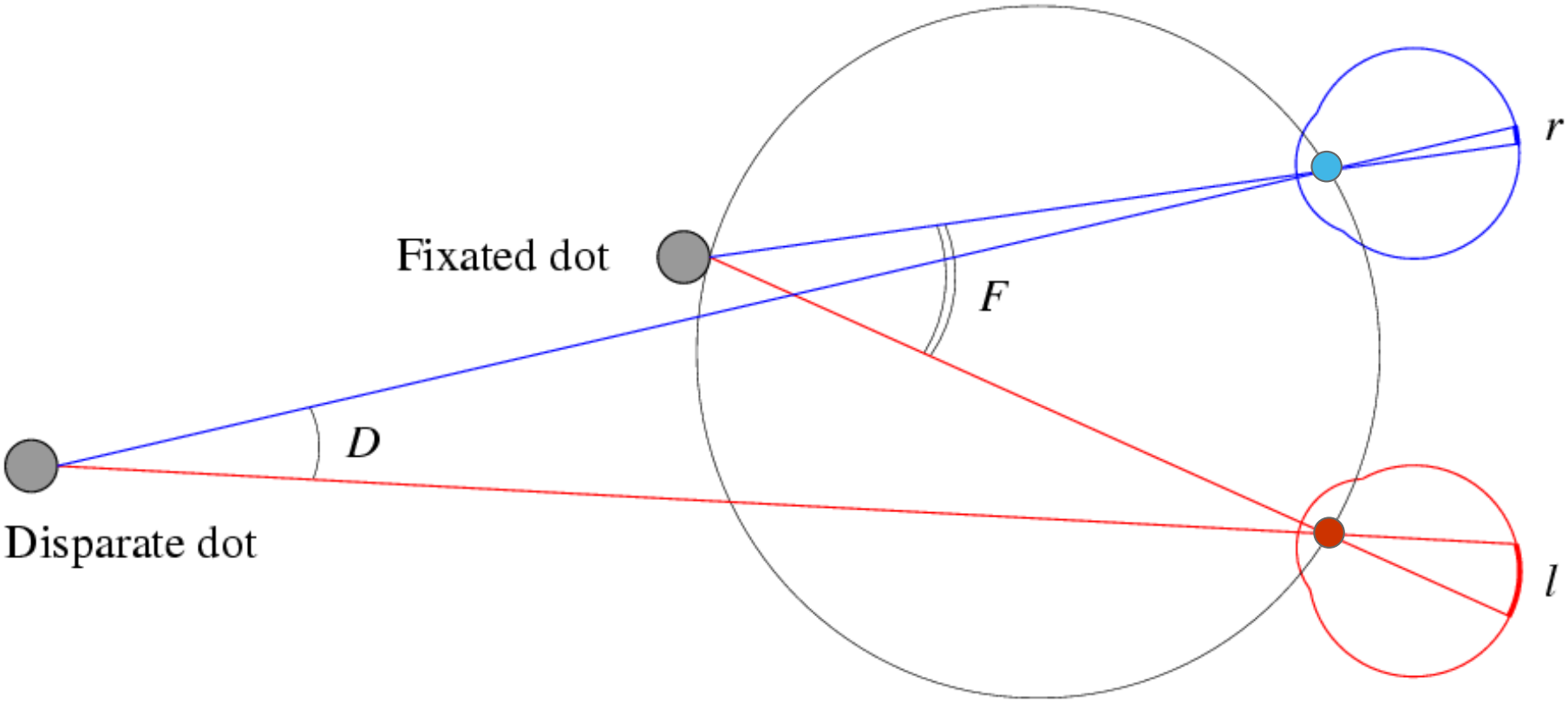
- **Disparity** occurs when eyes fixate on one object; others appear at different visual angles

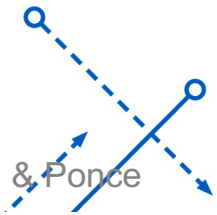


Disparity

- Disparity:

$$d = r - l = D - F.$$

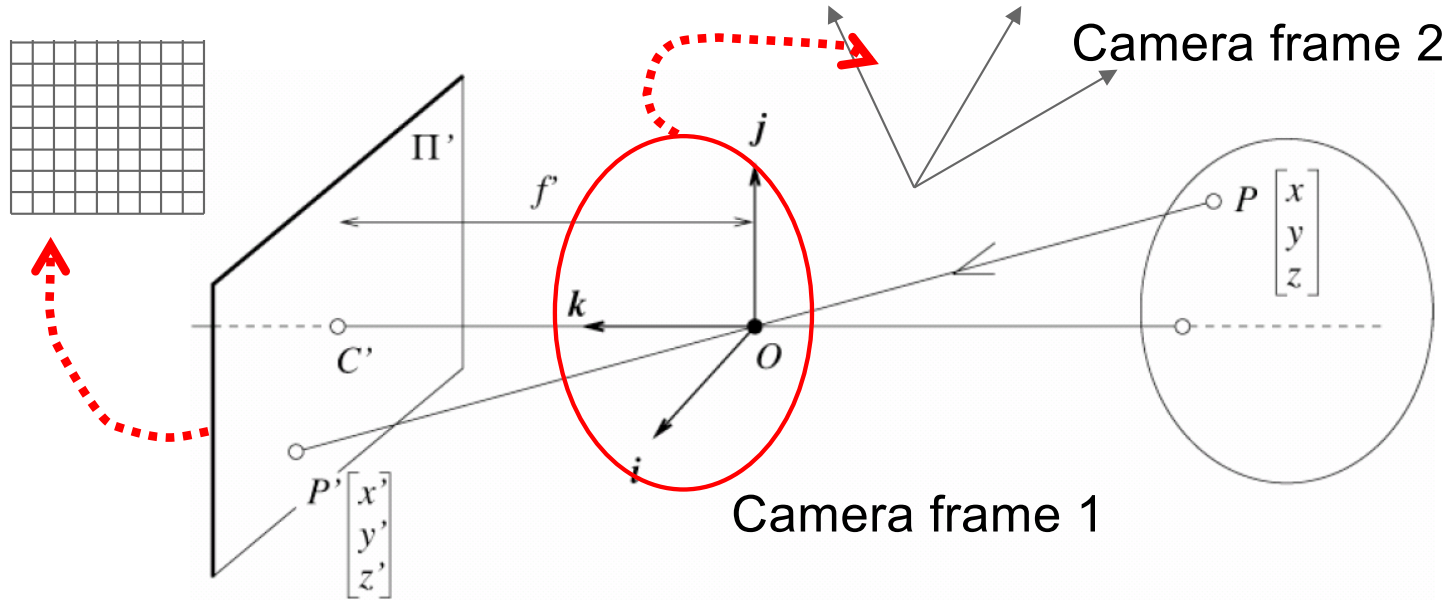




Forsyth & Ponce

Camera parameters (Recap)

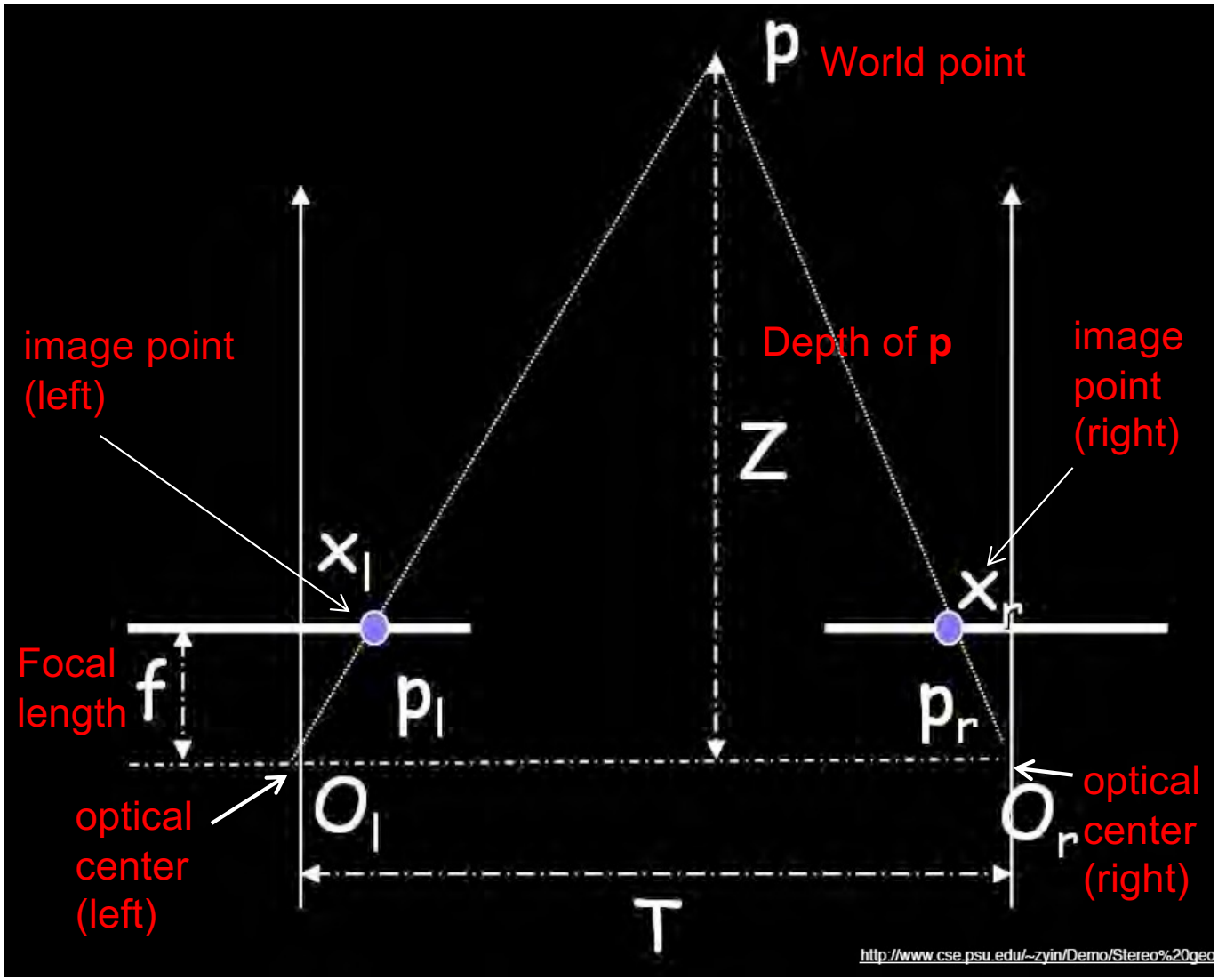
- *Extrinsic* params:
  - Rotation matrix and translation vector
  - Camera frame 1 ←→ Camera frame 2
- *Intrinsic* params:
  - Focal length, image center, radial distortion parameters
  - Coordinates relative to camera ←→ Pixel coordinates



We'll assume for now that these parameters are given and fixed.

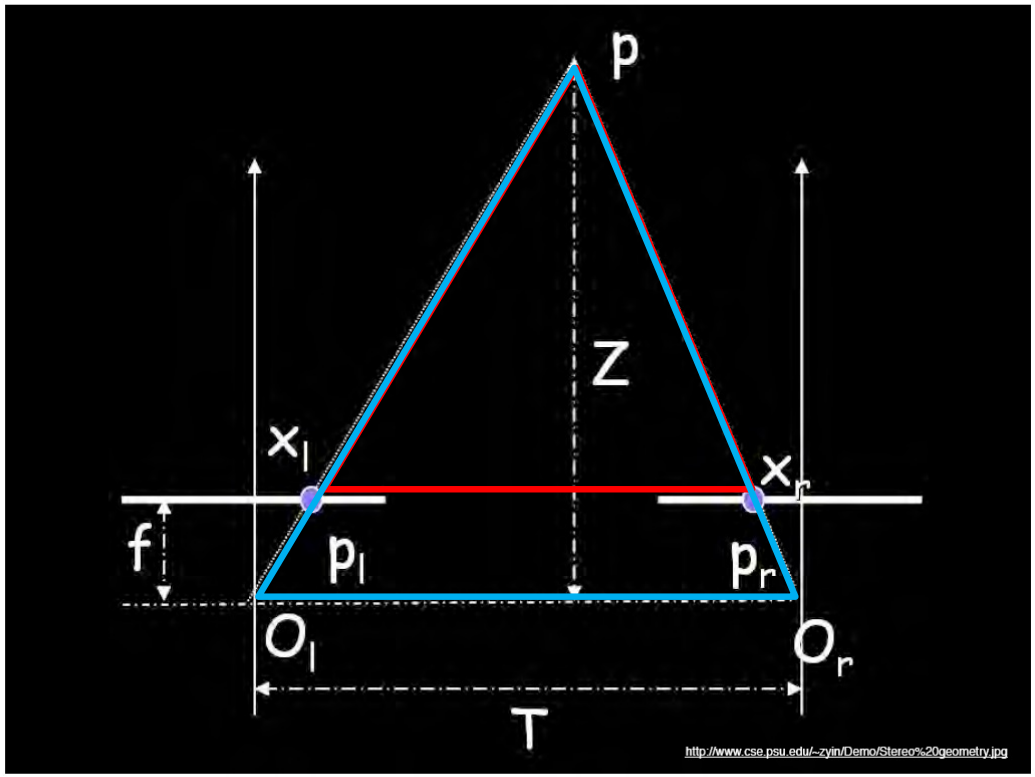
Geometry for a simple stereo system

- Assuming parallel optical axes, known camera parameters (i.e., calibrated cameras):



Geometry for a simple stereo system

- Assume parallel optical axes, known camera parameters, i.e., calibrated cameras. **What is expression for Z?**
- Similar triangles (p_l, P, p_r) and (O_l, P, O_r):



$$\frac{T - x_l + x_r}{Z - f} = \frac{T}{Z}$$

$$Z = f\frac{T}{x_l - x_r}$$

Disparity

Depth from disparity

- If we could find the **corresponding points** in two images, we could **estimate relative depth.**

Image $I(x, y)$

Disparity map $D(x, y)$

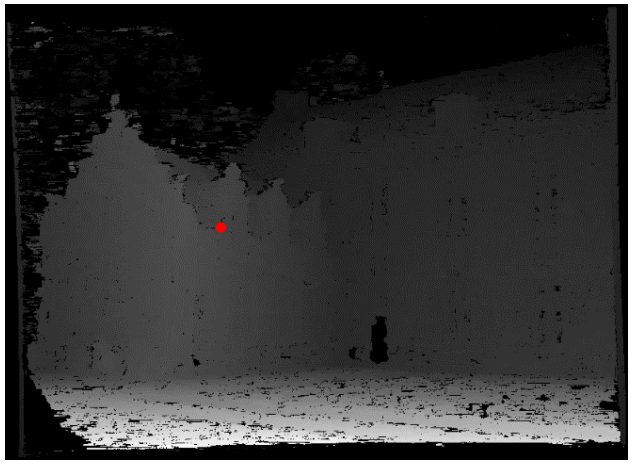
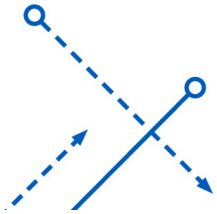


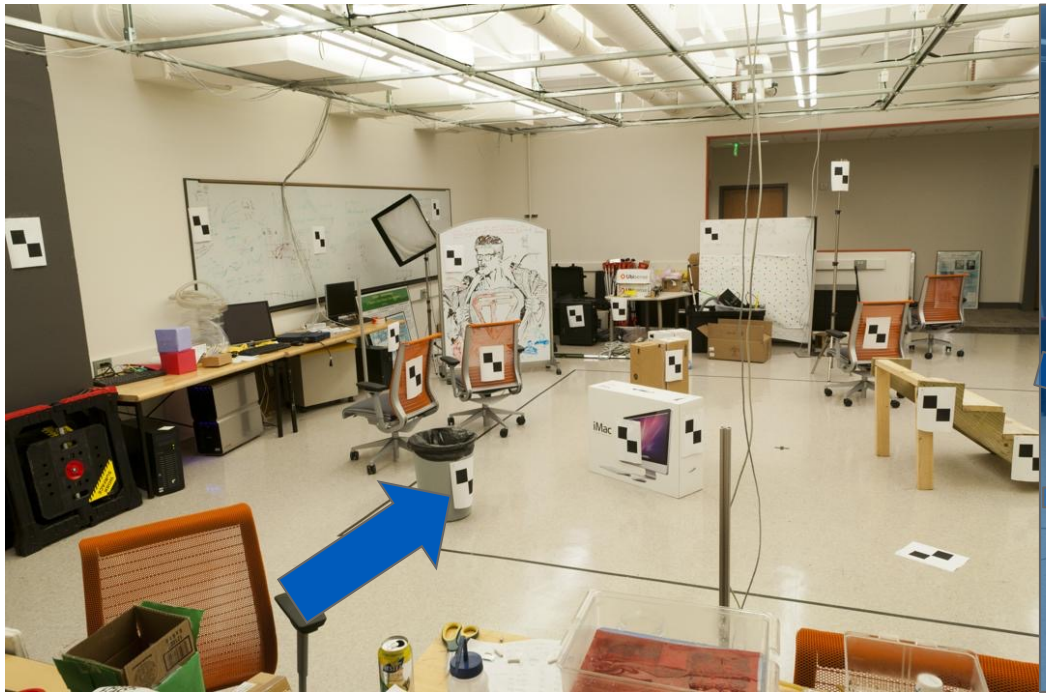
Image $I'(x', y')$

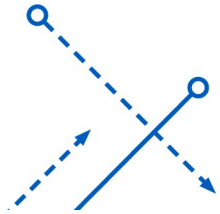
$$(x', y') = (x + D(x, y), y)$$



Depth from disparity

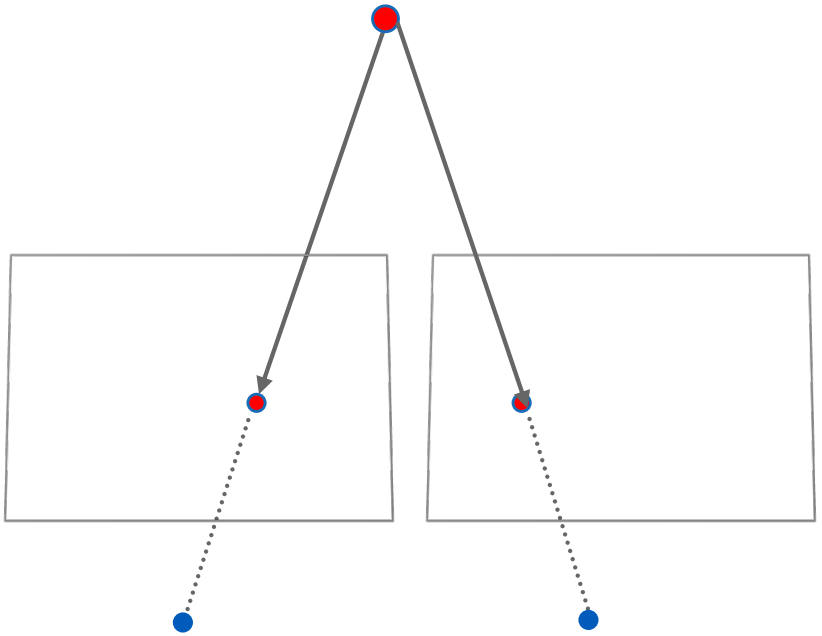
- Where do we need to search?



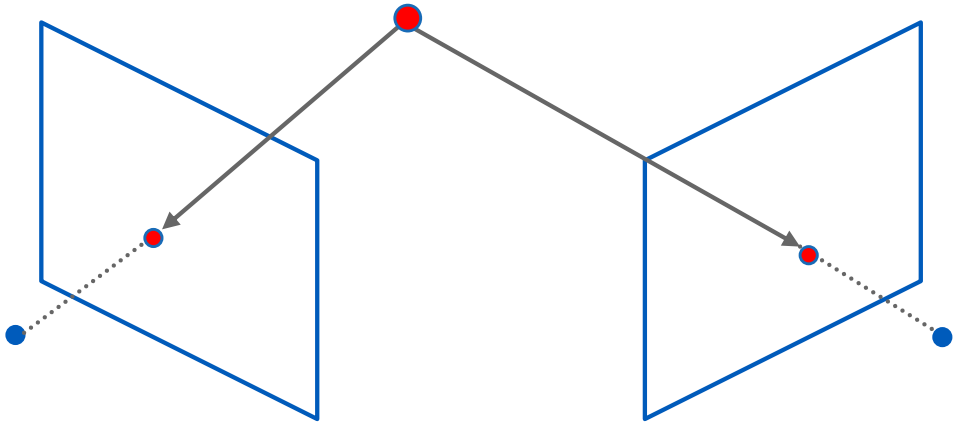


General case (calibrated cameras)

- The two cameras need not have parallel optical axes.

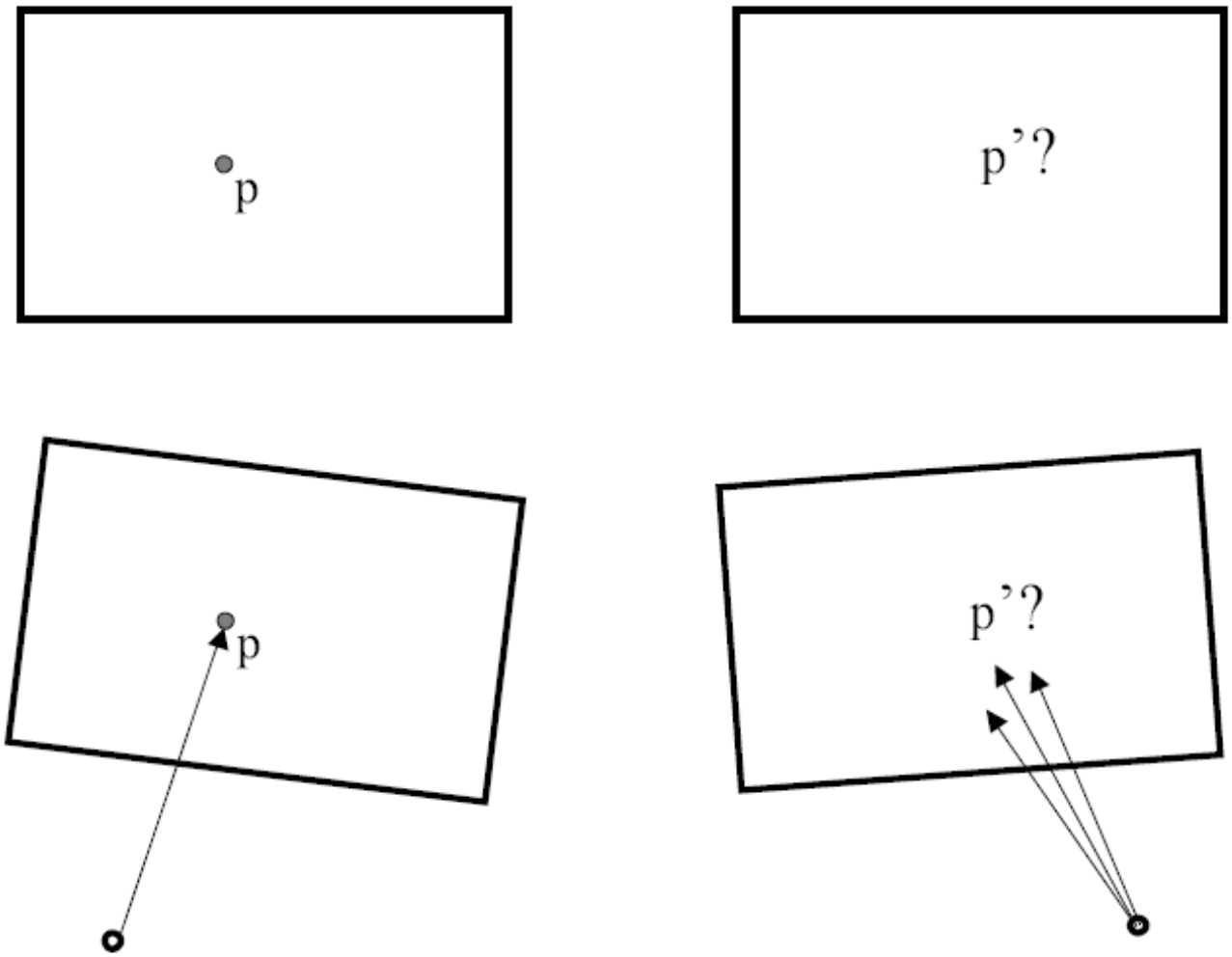


Vs.



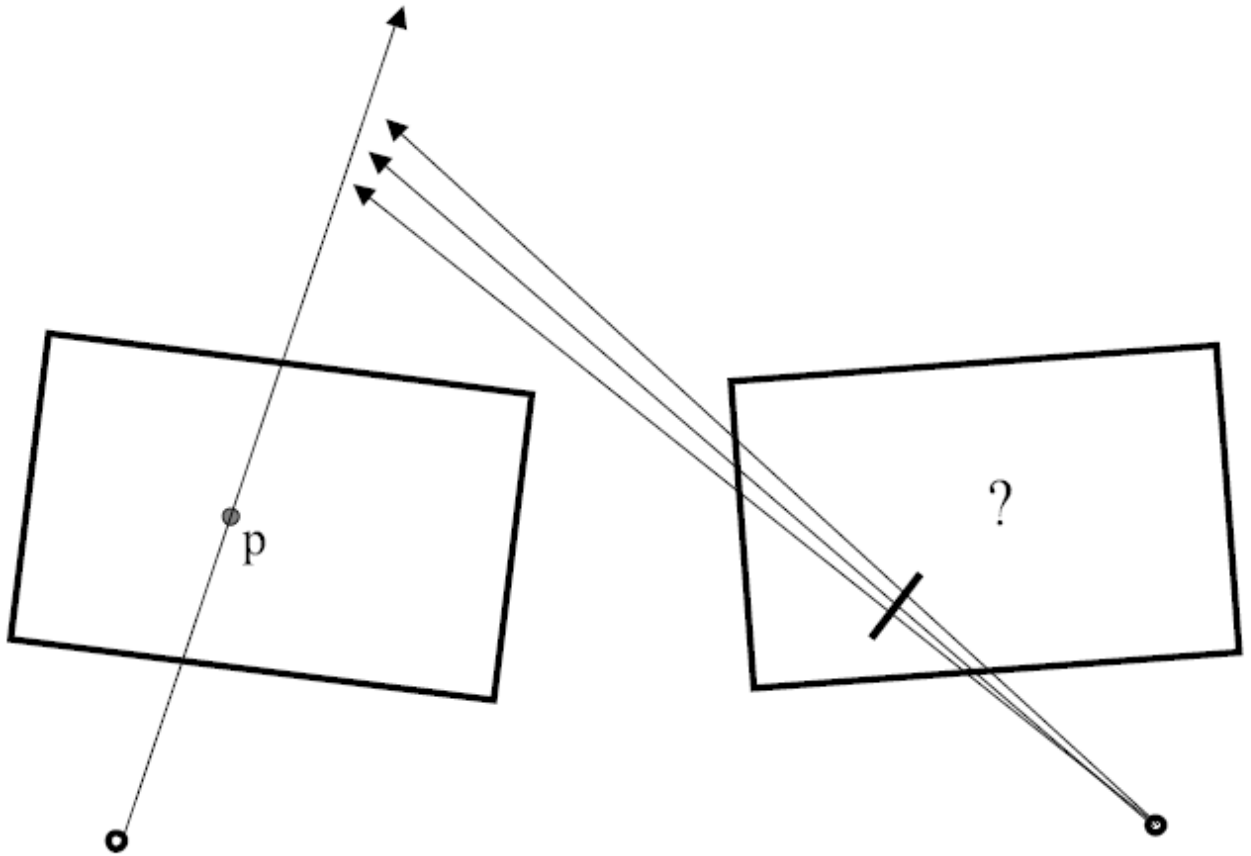
Stereo correspondence constraints

- Given p in left, where can corresponding point p' be?



Stereo correspondence constraints

- Given p in left, where can corresponding point p' be?



Correspondence problem

- Multiple match hypotheses satisfy epipolar constraint, but which is correct?

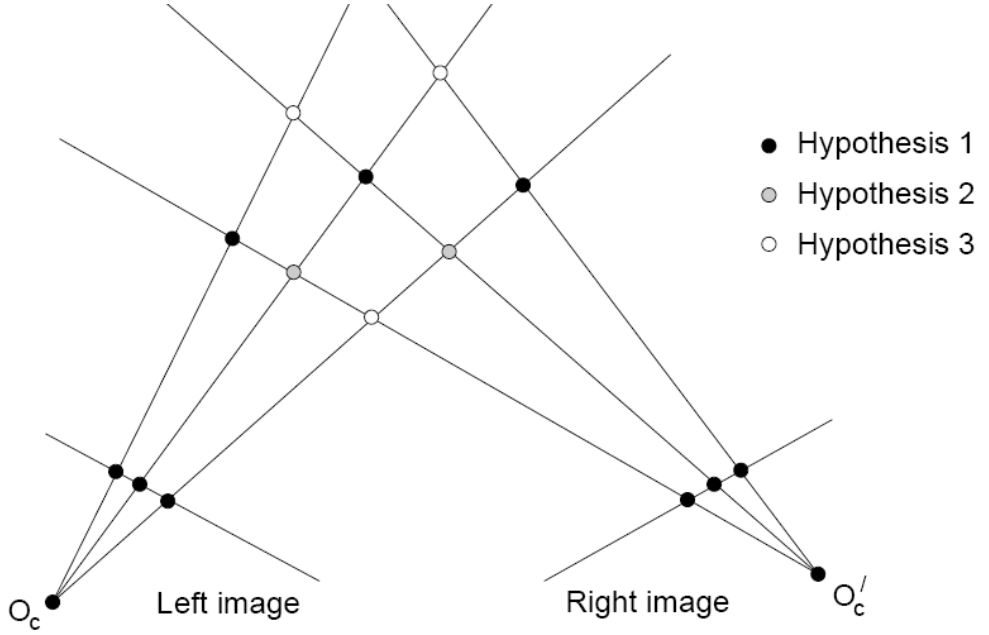
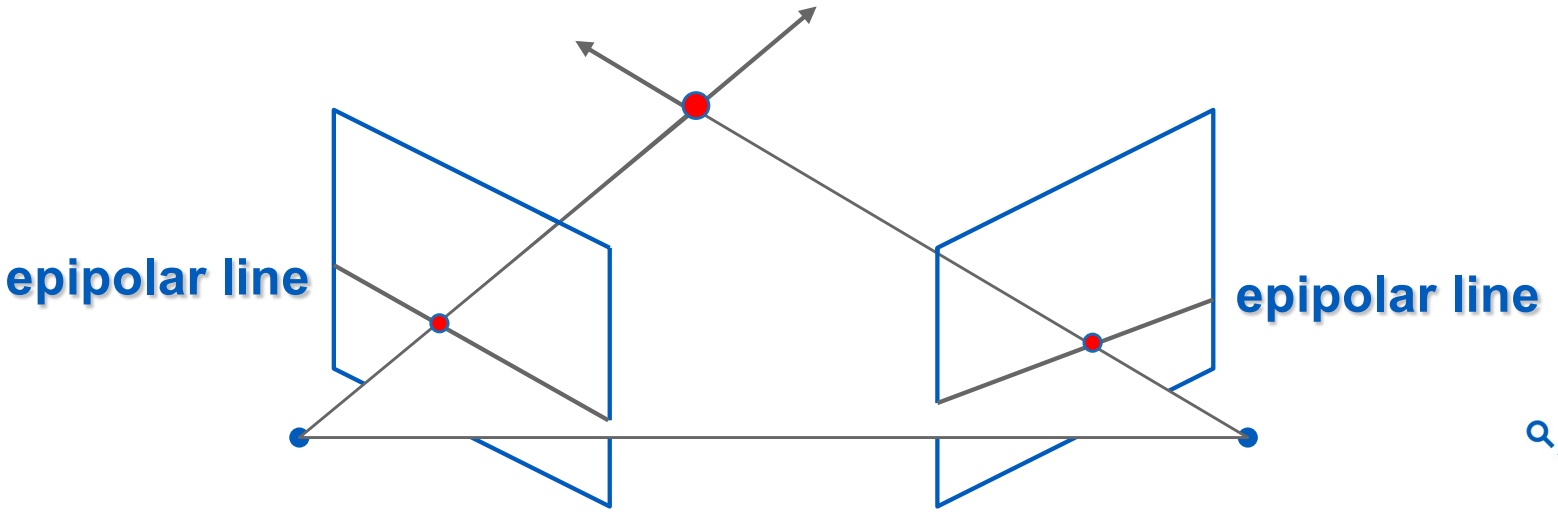


Figure from Gee & Cipolla 1999

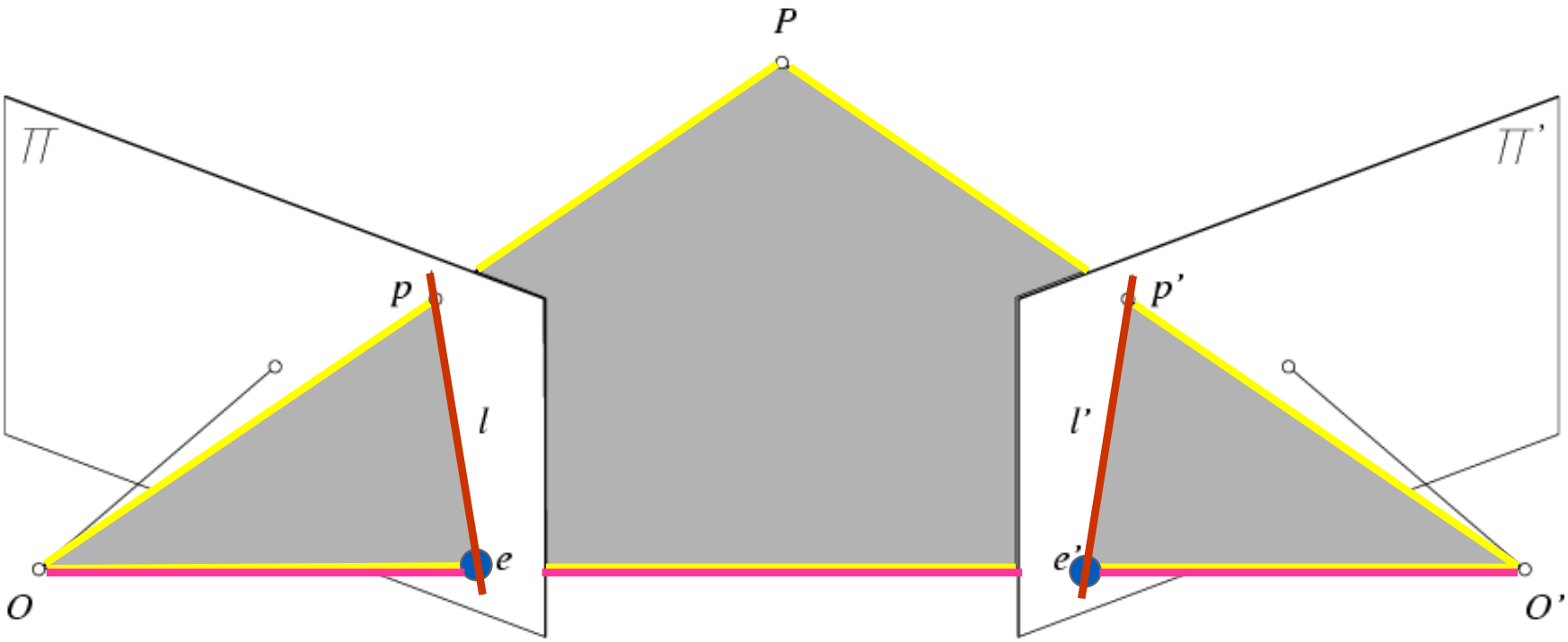
Stereo correspondence constraints

- Geometry of two views allows us to constrain where the corresponding pixel for image points in the first view must occur in the second view.
- Epipolar constraint
  - Reduces correspondence problem to 1D search along conjugate epipolar lines

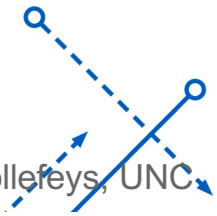


Adapted from Steve Seitz

Epipolar geometry

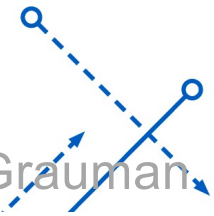


- Epipolar Plane
- Baseline
- Epipoles
- Epipolar Lines



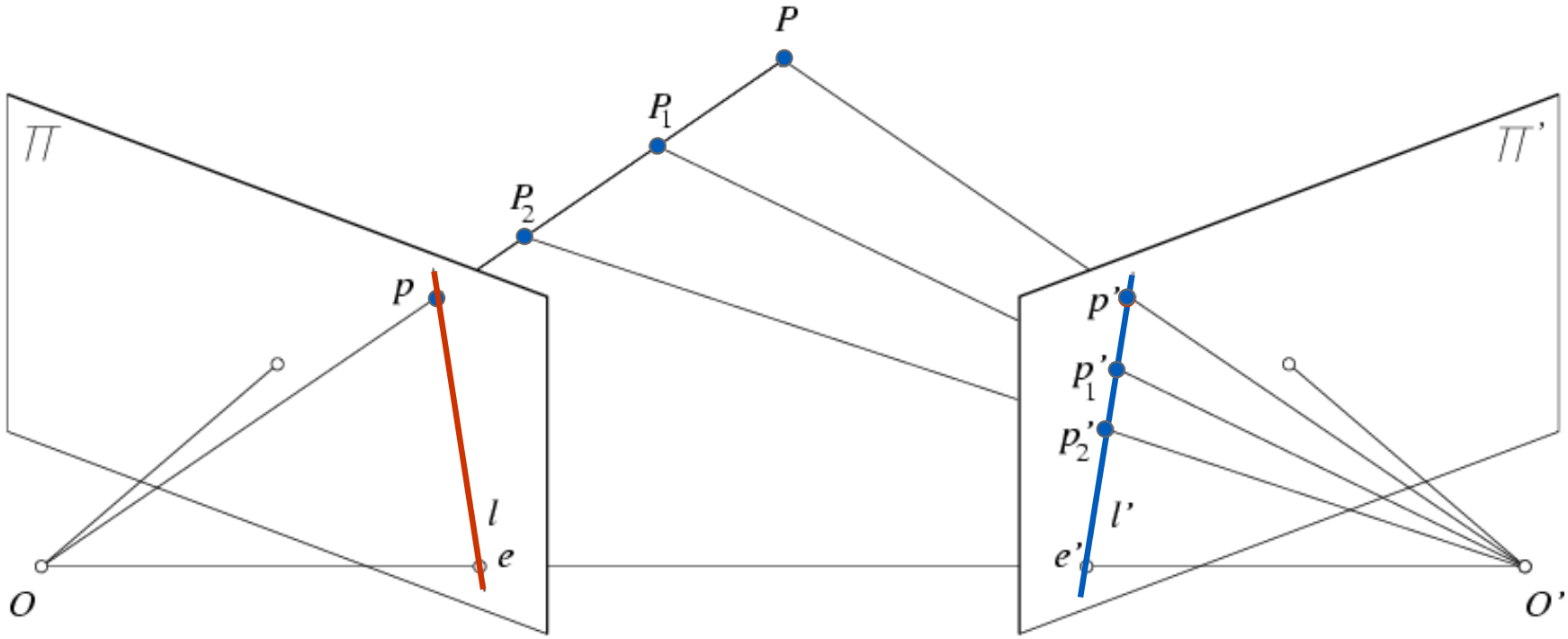
Epipolar geometry: terms

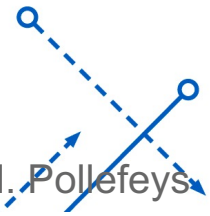
- **Baseline**: line joining the camera centers
- **Epipole**: point of intersection of baseline with the image plane
- **Epipolar plane**: plane containing baseline and world point
- **Epipolar line**: intersection of epipolar plane with image plane
- All epipolar lines intersect at the epipole
- An epipolar plane intersects the left and right image planes in epipolar lines



Epipolar constraint

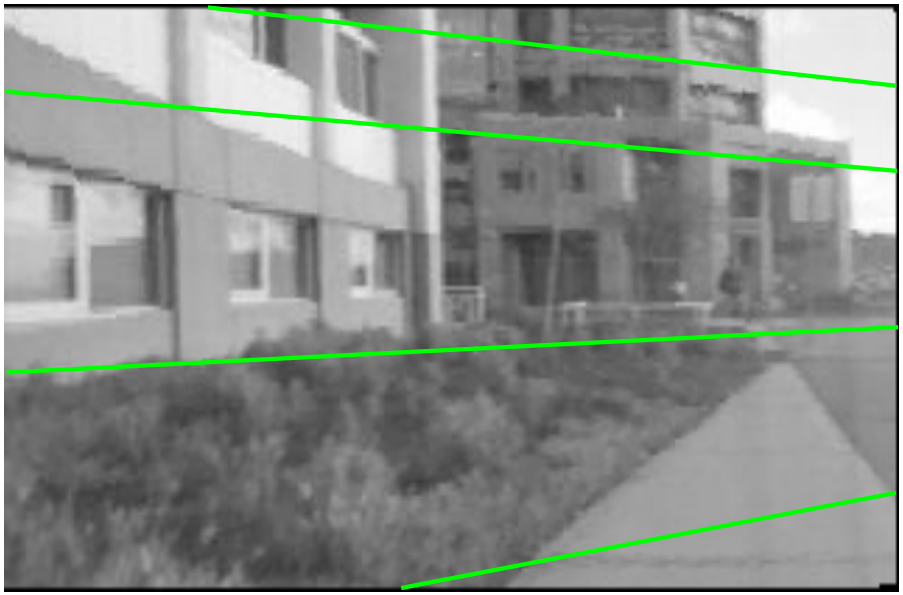
- Potential matches for p must lie on epipolar line l'.
- Potential matches for p' must lie on epipolar line l.

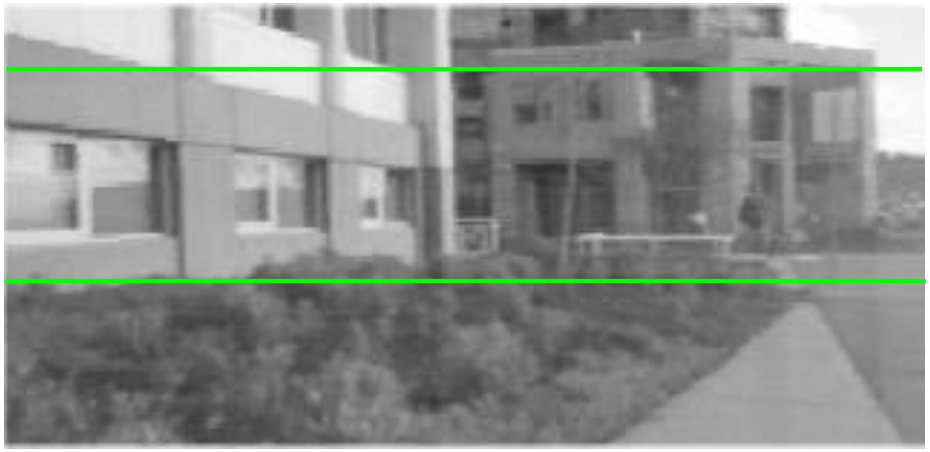


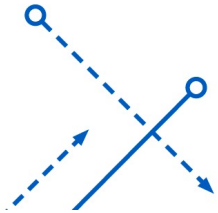


Rectification

- Searching along epipolar lines at arbitrary orientation is intuitively expensive.
- We prefer searching along the image row.
  - Epipolar lines parallel to the rows of the image.
- This transformation is called *rectification*.

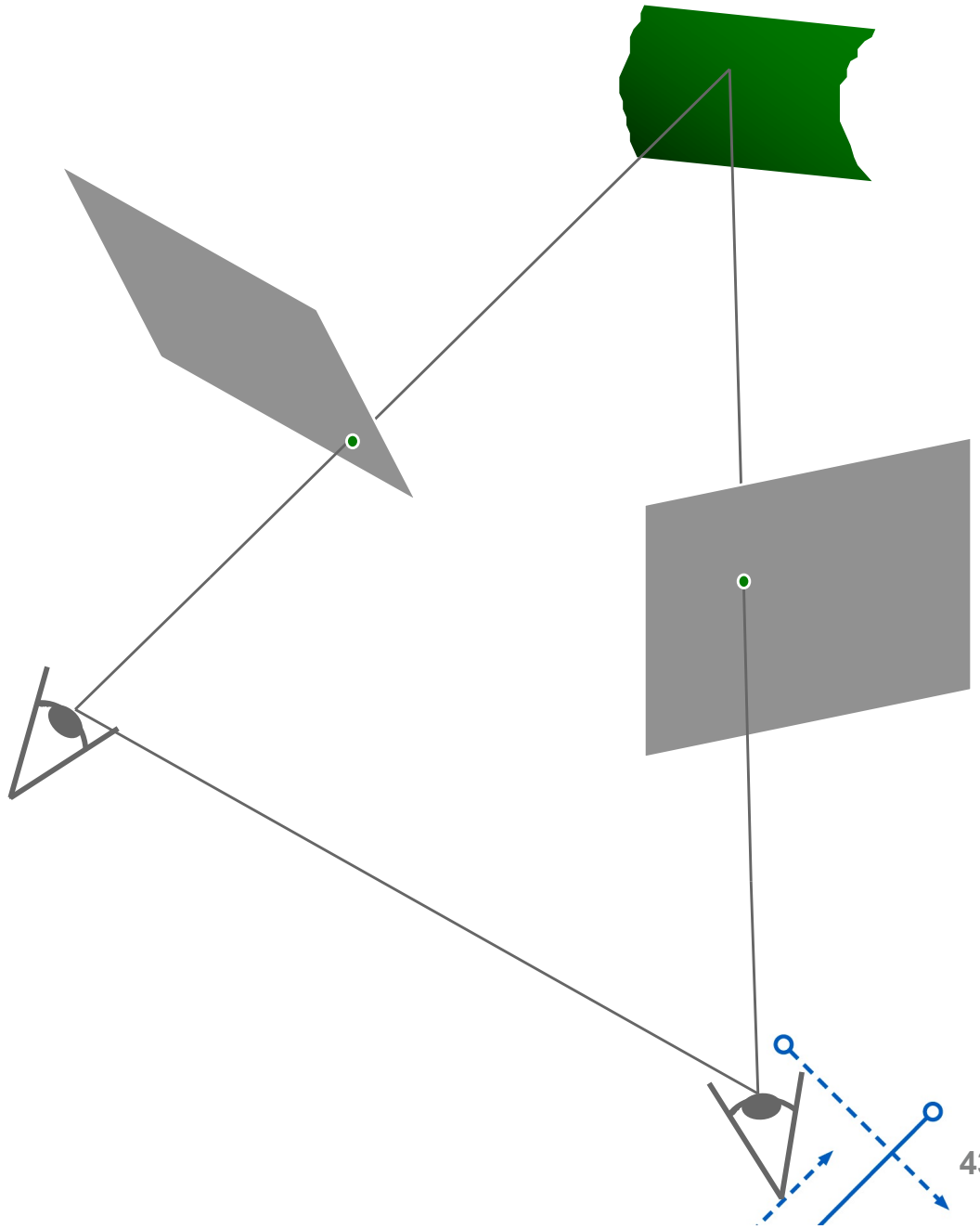






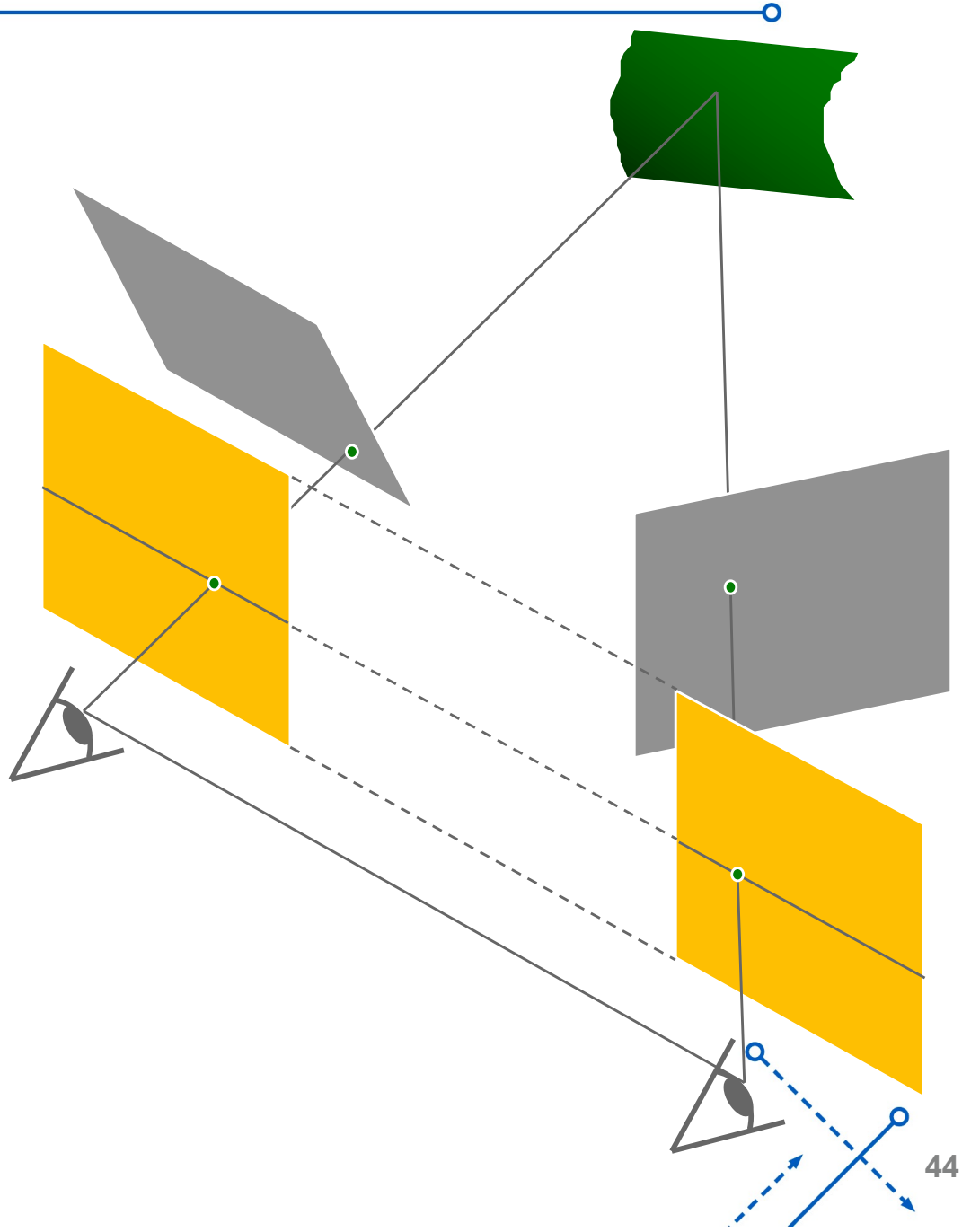
Stereo Image Rectification

- Reproject image planes onto a common plane parallel to the line between optical centers

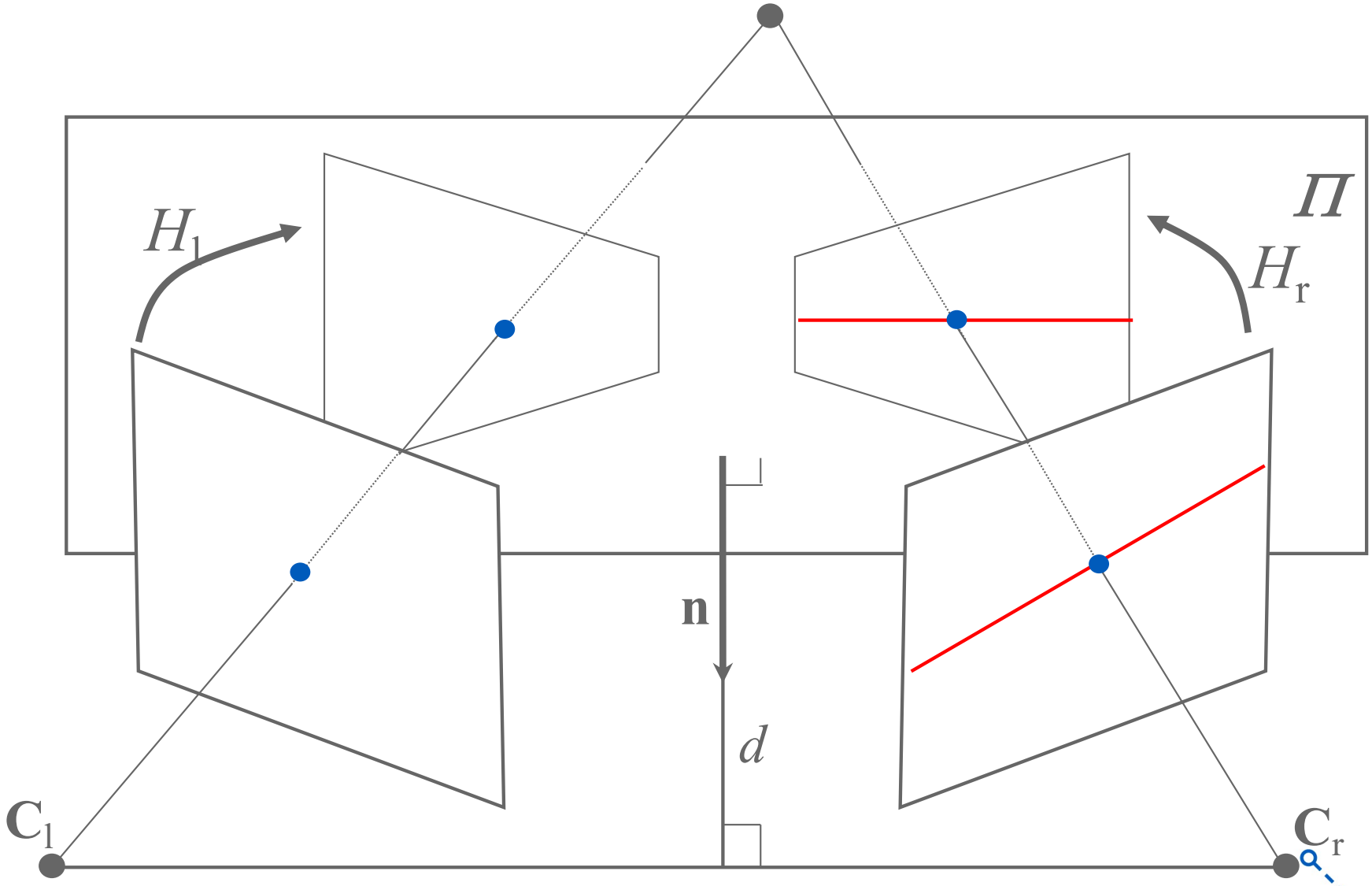


Stereo Image Rectification

- Reproject image planes onto a common plane parallel to the line between optical centers
- pixel motion is horizontal after this transformation



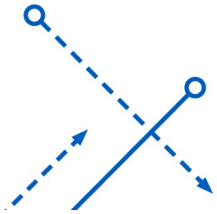
Stereo Image Rectification



Correspondence Search

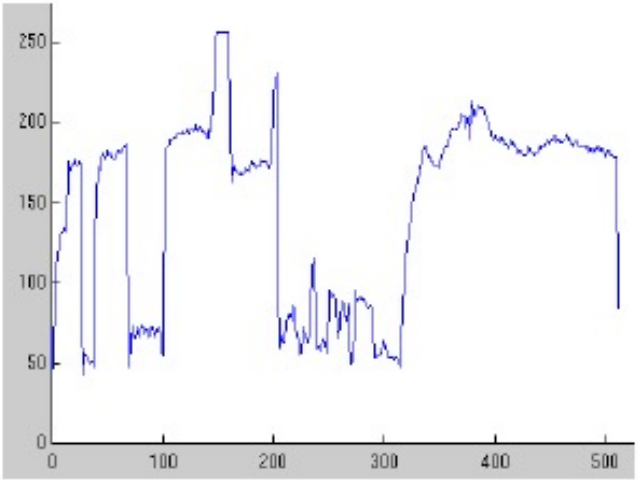
- Other “soft” constraints (To cover)
  - 1. Similarity
  - 2. Uniqueness
  - 3. Disparity gradient
  - 4. Ordering

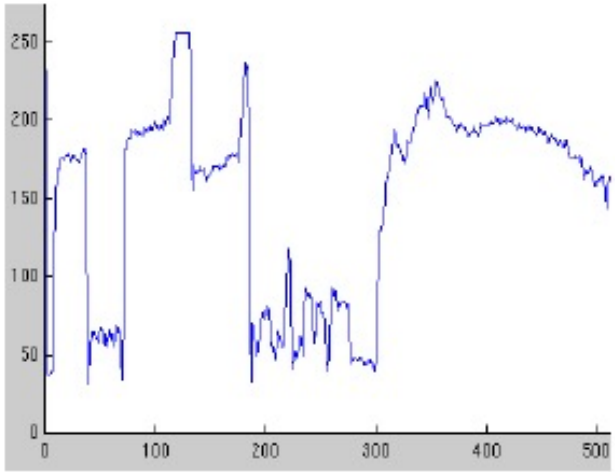
- To find matches in the image pair, we will assume
  - Most scene points visible from both views
  - Matched regions are similar in appearance



Intensity profiles

Intensity profiles



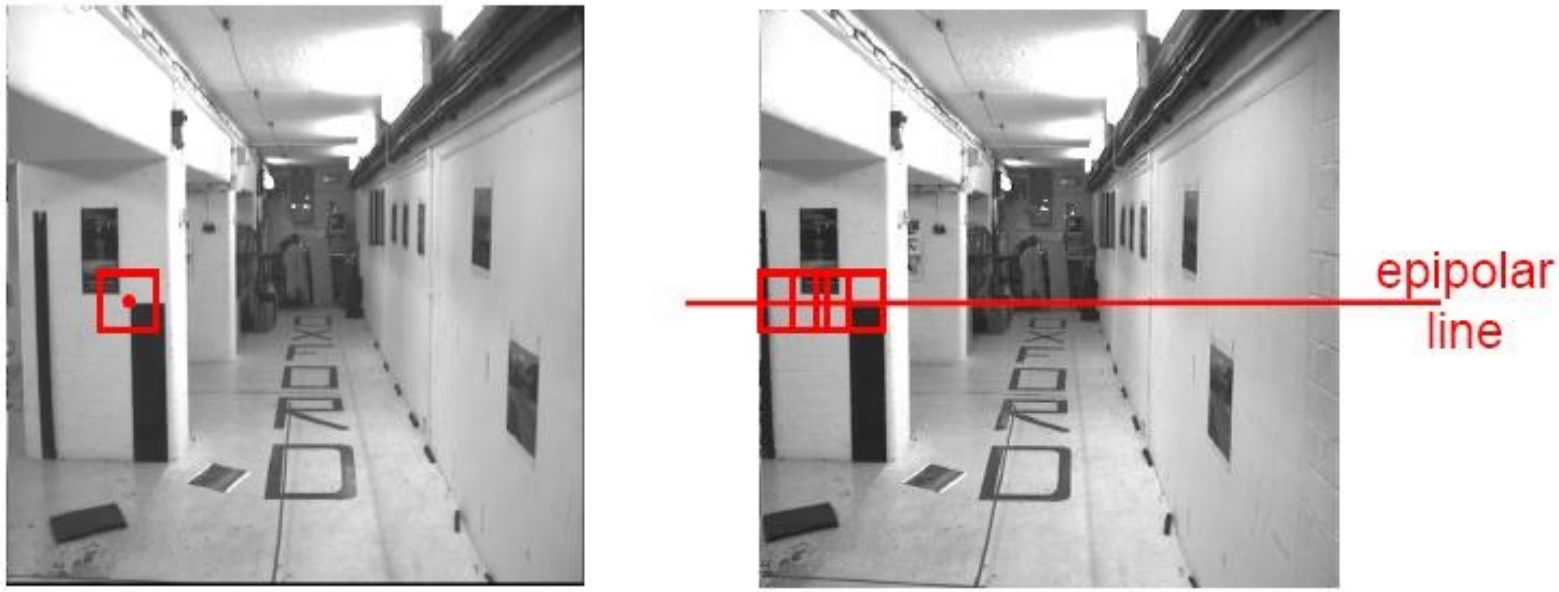


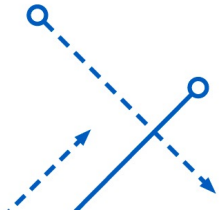
- Clear correspondence between intensities, but also noise and ambiguity

Source: Andrew Zisserman

Dense correspondence search

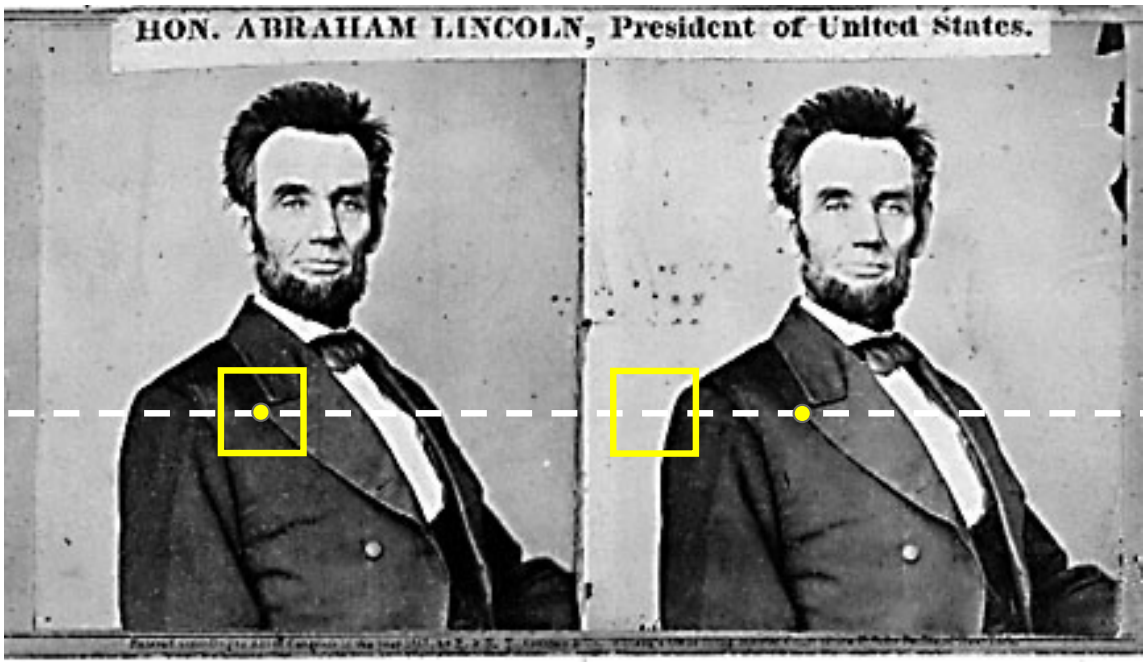
- Neighborhoods of corresponding points are similar in intensity patterns.





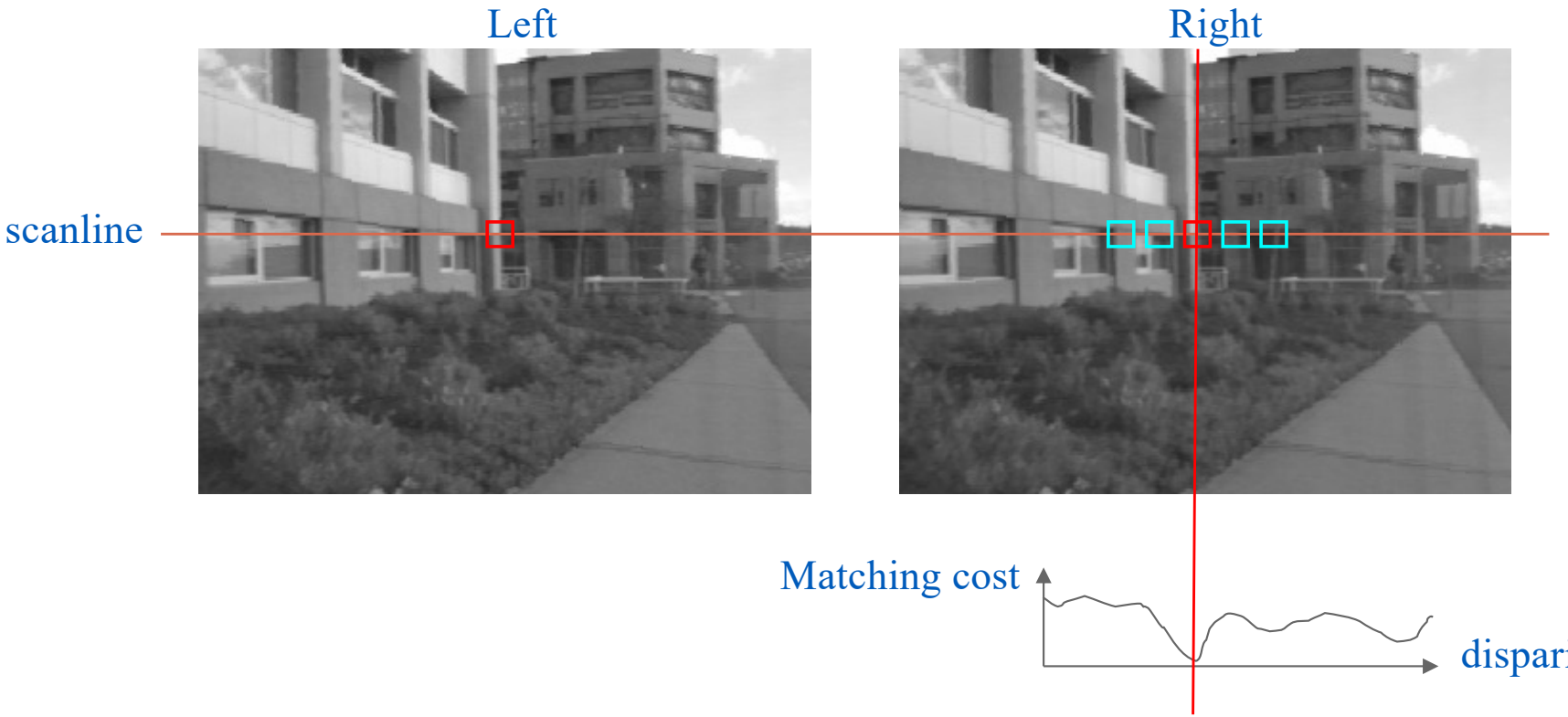
Dense correspondence search

- For each epipolar line
  - For each pixel / window in the left image
    - Compare with every pixel / window on same epipolar line
    - Pick position with minimum match cost
      - SSD, normalized correlation

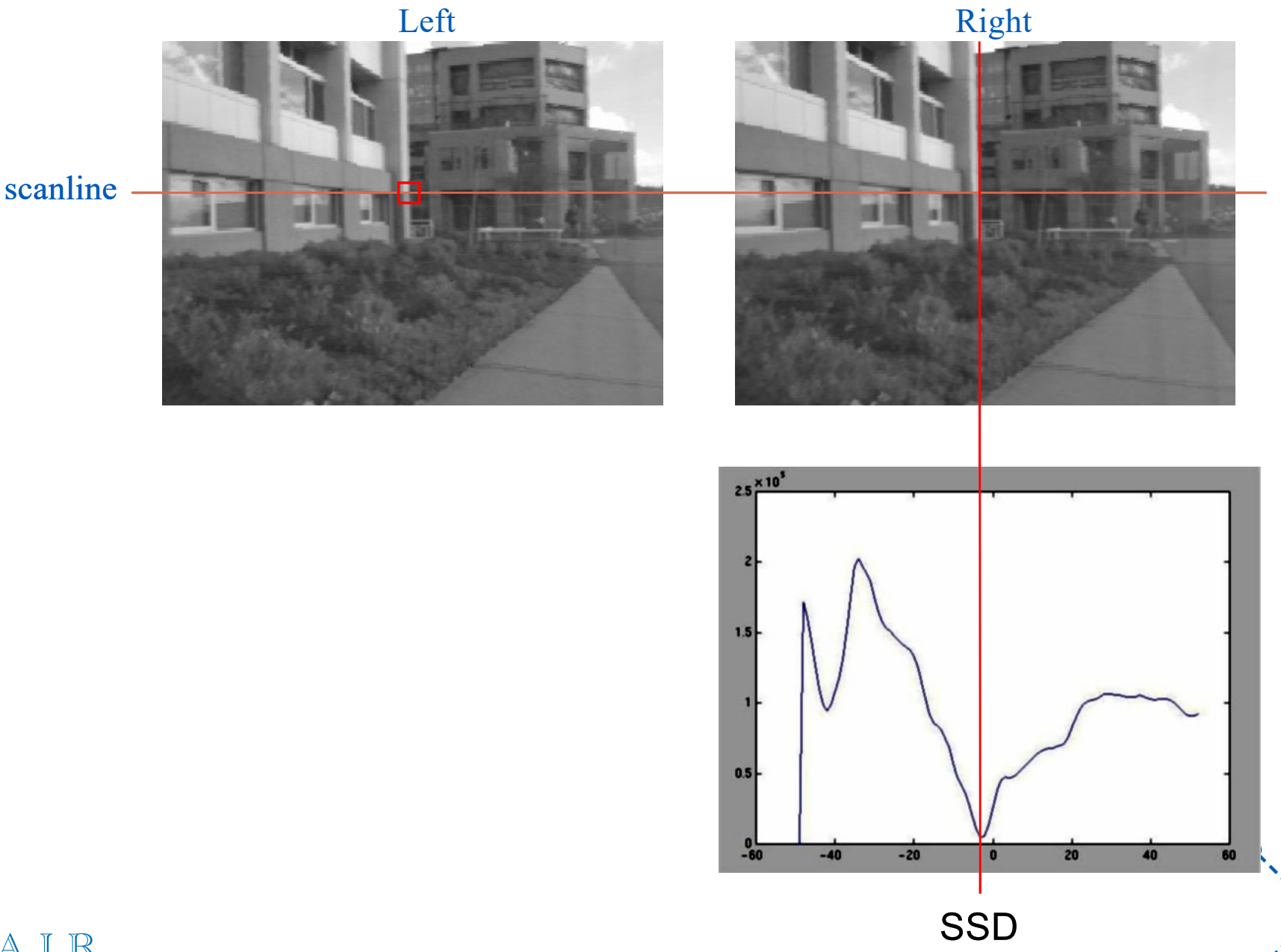


Similarity constraints

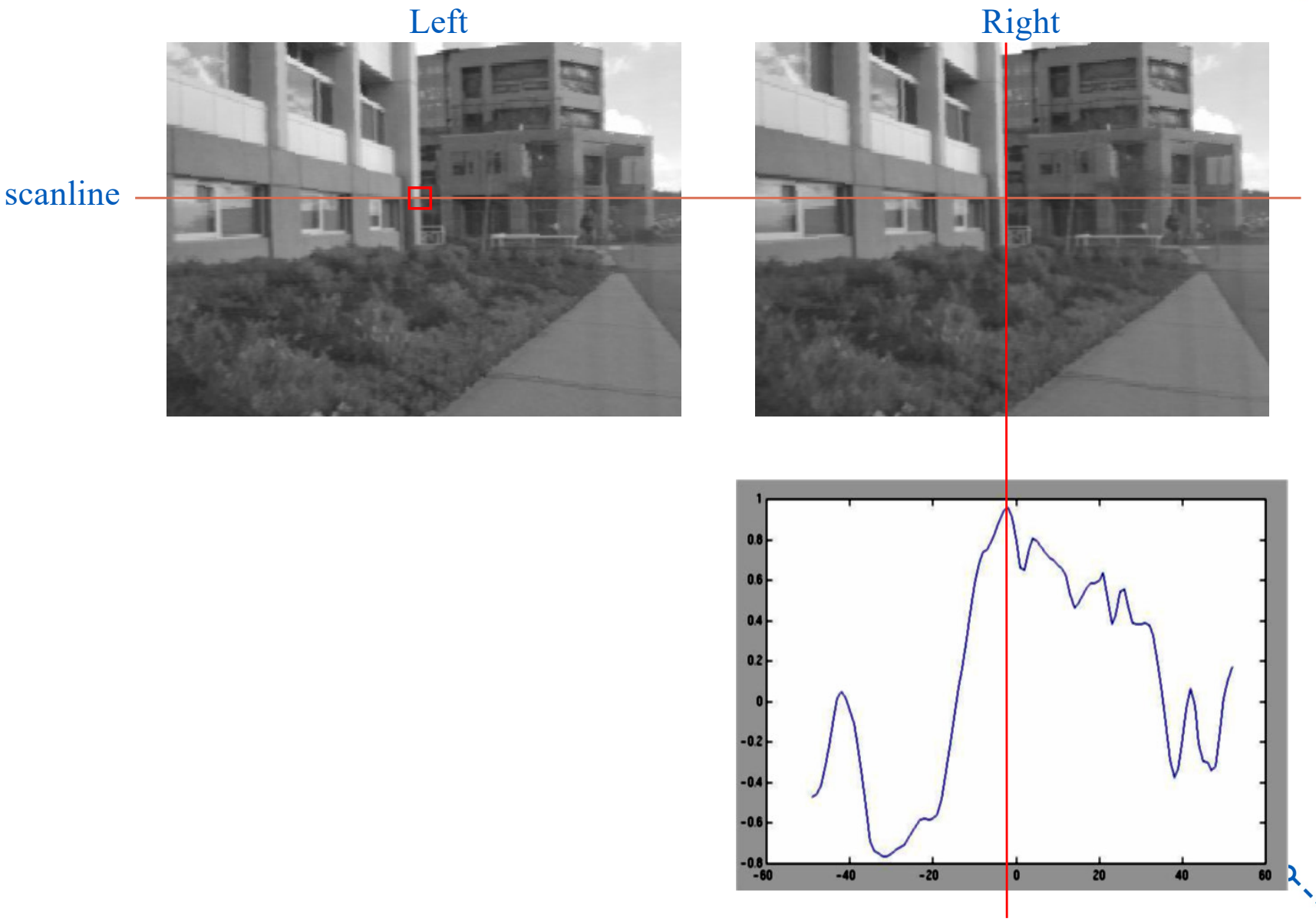
- Slide a window along the right scanline and compare contents of that window with the reference window in the left image
- Matching cost: SSD or normalized correlation



Similarity constraints: SSD



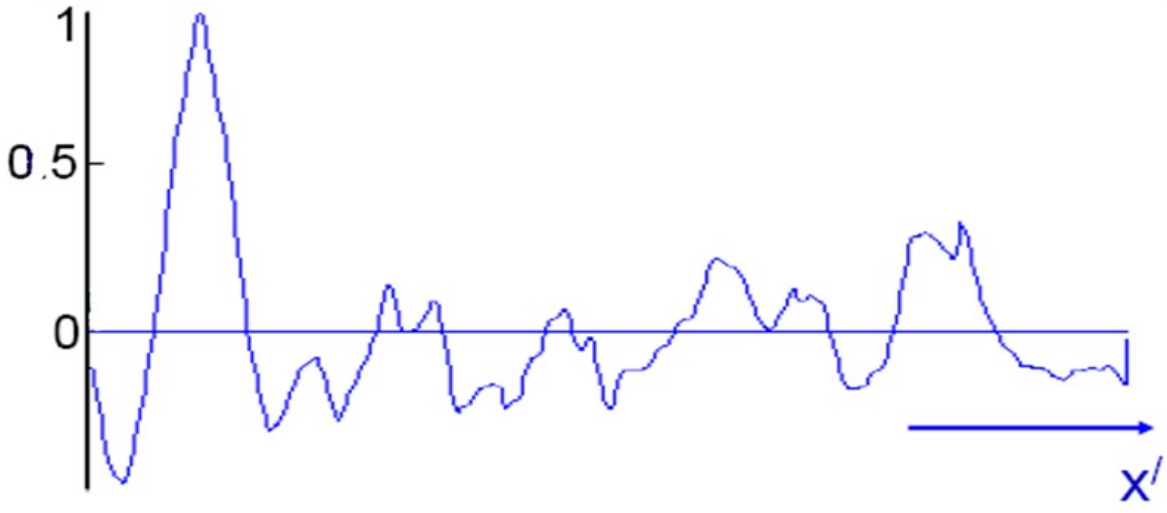
Similarity constraints: Norm. Corr.



Correlation-based window matching

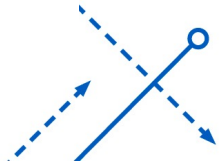
left image band (x)

right image band ($x^{/}$)

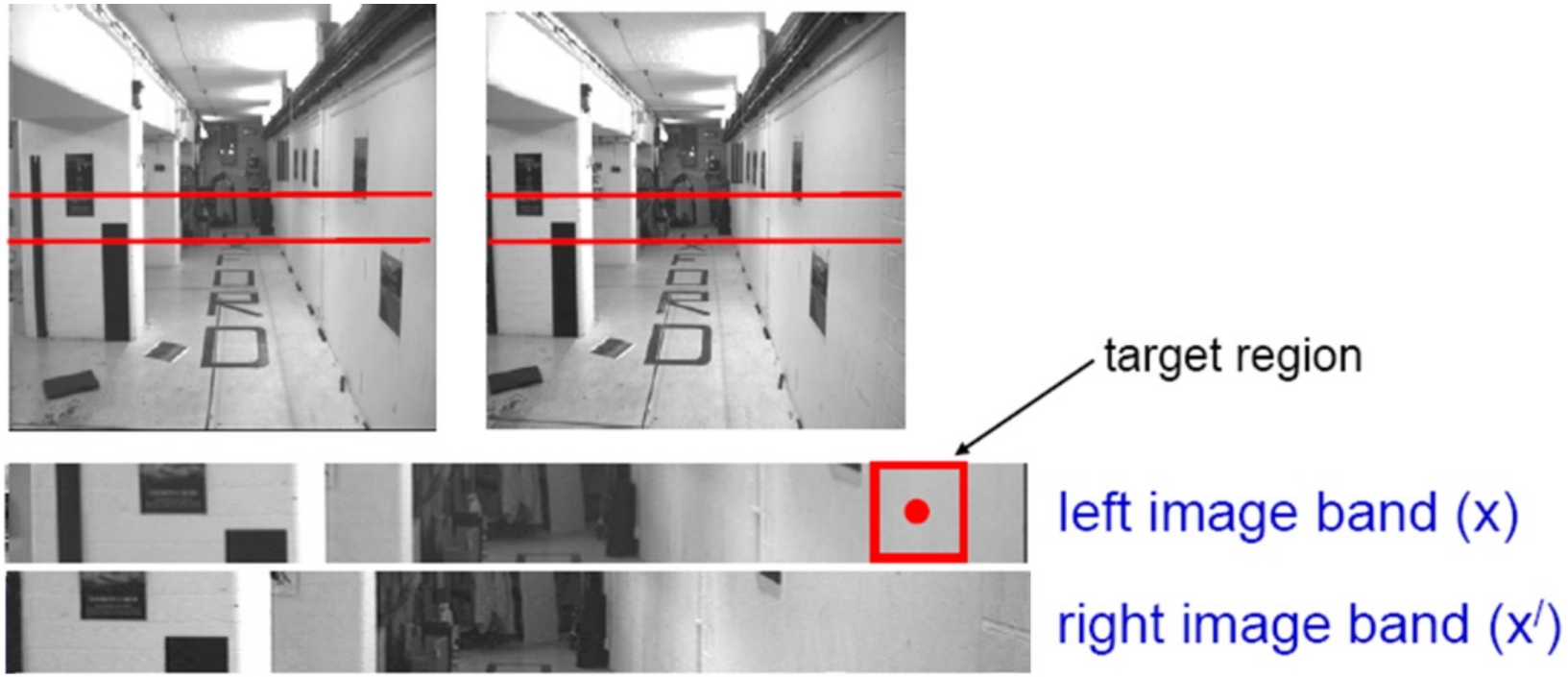


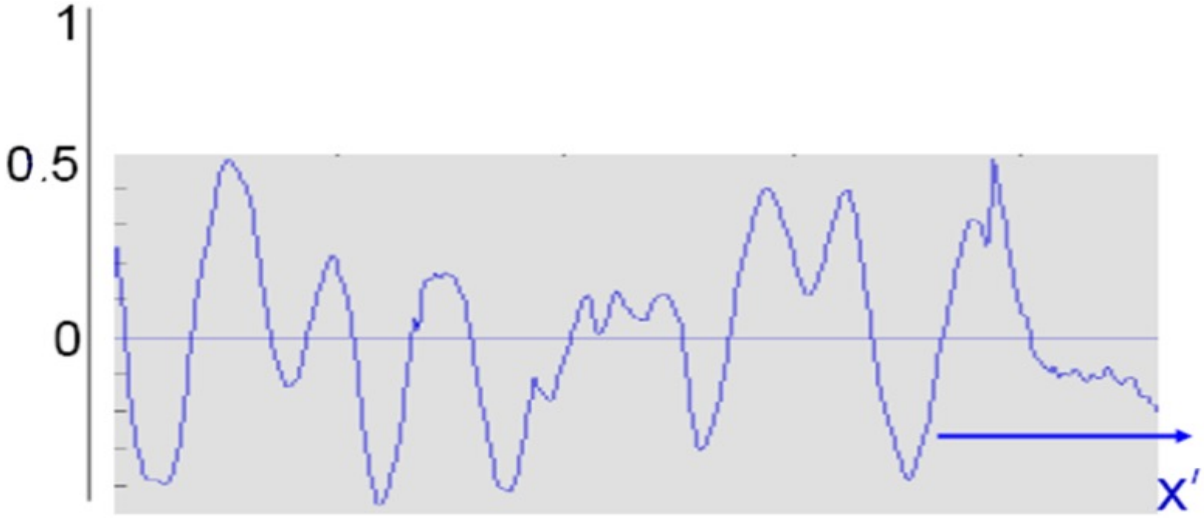
cross correlation

disparity = $x^{/}$ - x



Correlation-based window matching



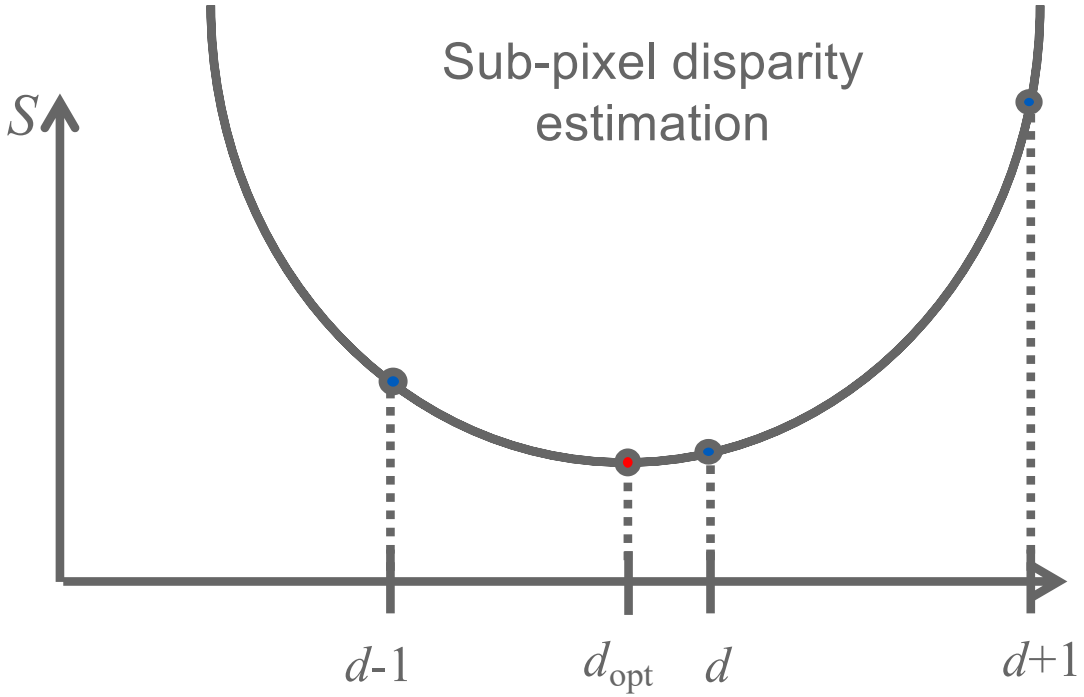


cross correlation

Textureless regions are non-distinct; high ambiguity for matches.

Sub-pixel disparity estimation

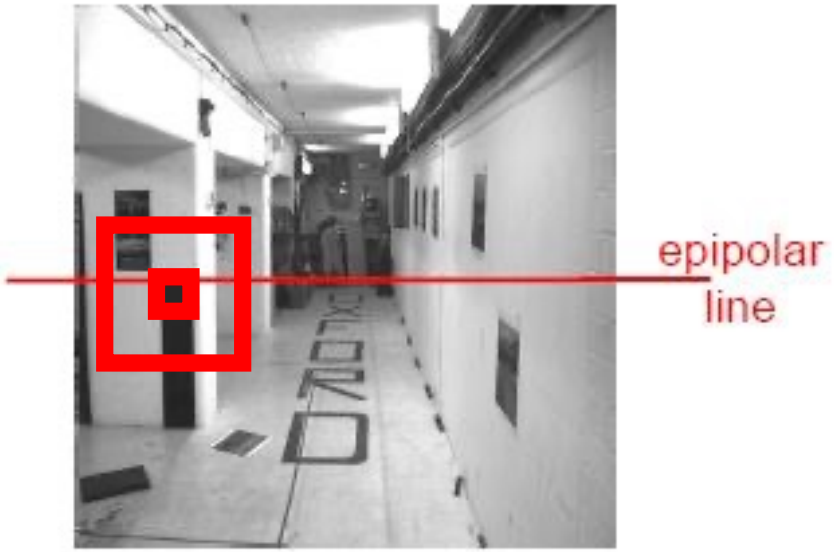
- Let S be the SSD



- $S(d) = ad^2 + bd + c$
  - $S(0) = c$
  - $S(1) = a + b + c$
  - $S(-1) = a - b + c$
- Solving this, we obtain:
  - $a = (S(1) + S(-1) - 2S(0))/2$
  - $b = (S(1) - S(-1))/2$
  - $c = S(0)$
- $S'(d) = 2ad + b = 0$

$$d_{opt} = \frac{(S(-1) - S(1))}{2(S(1) + S(-1) - 2S(0))}$$

Effect of window size

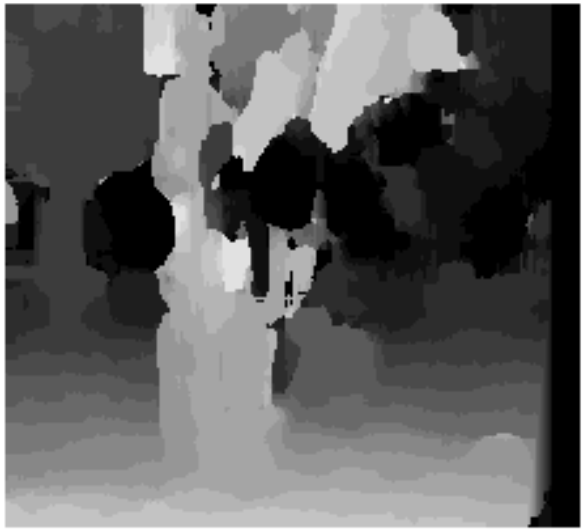


Source: Andrew Zisserman

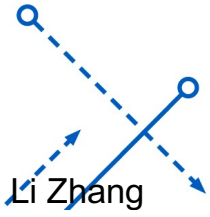
Effect of window size

- large enough to have sufficient intensity variation
- small enough to contain only pixels with about the same disparity.

W = 3



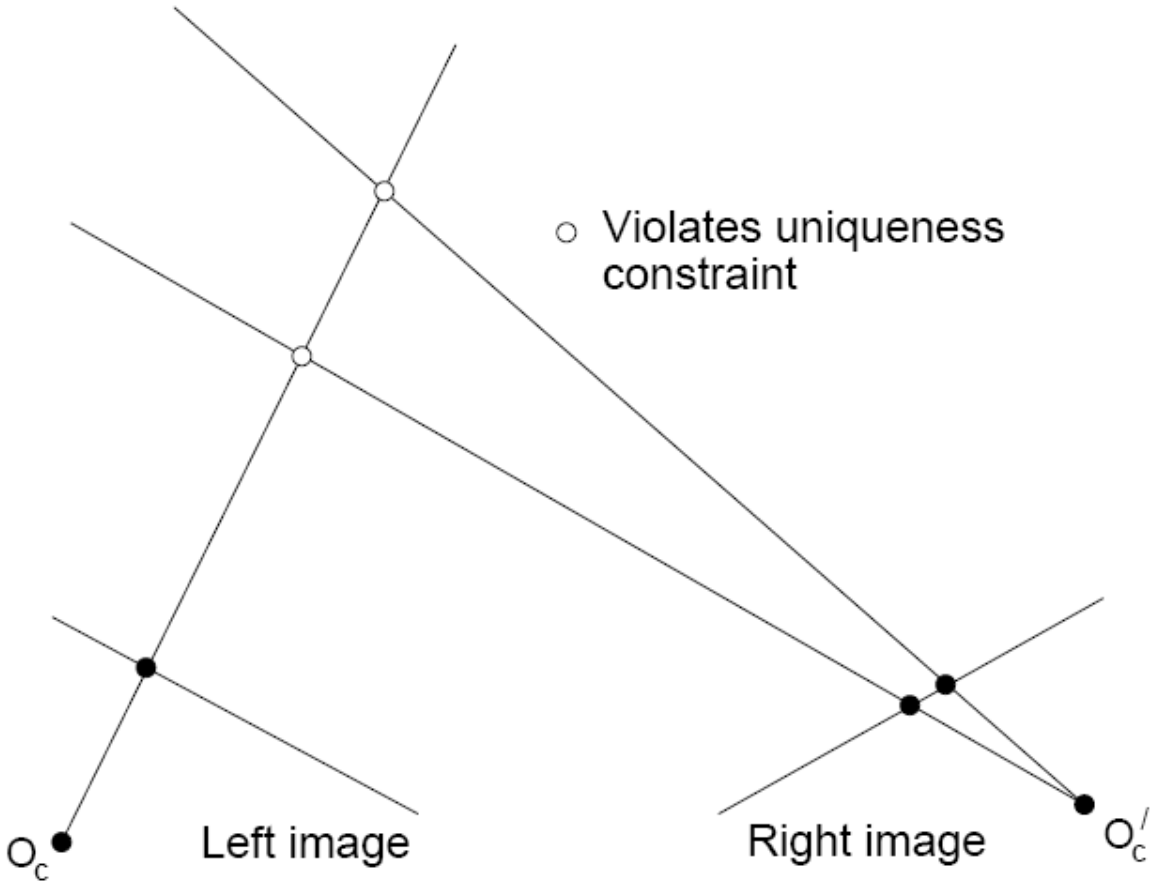
W = 20



from Li Zhang

Uniqueness constraint

- Up to one match in right image for every point in left image



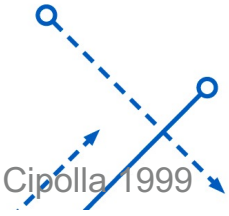
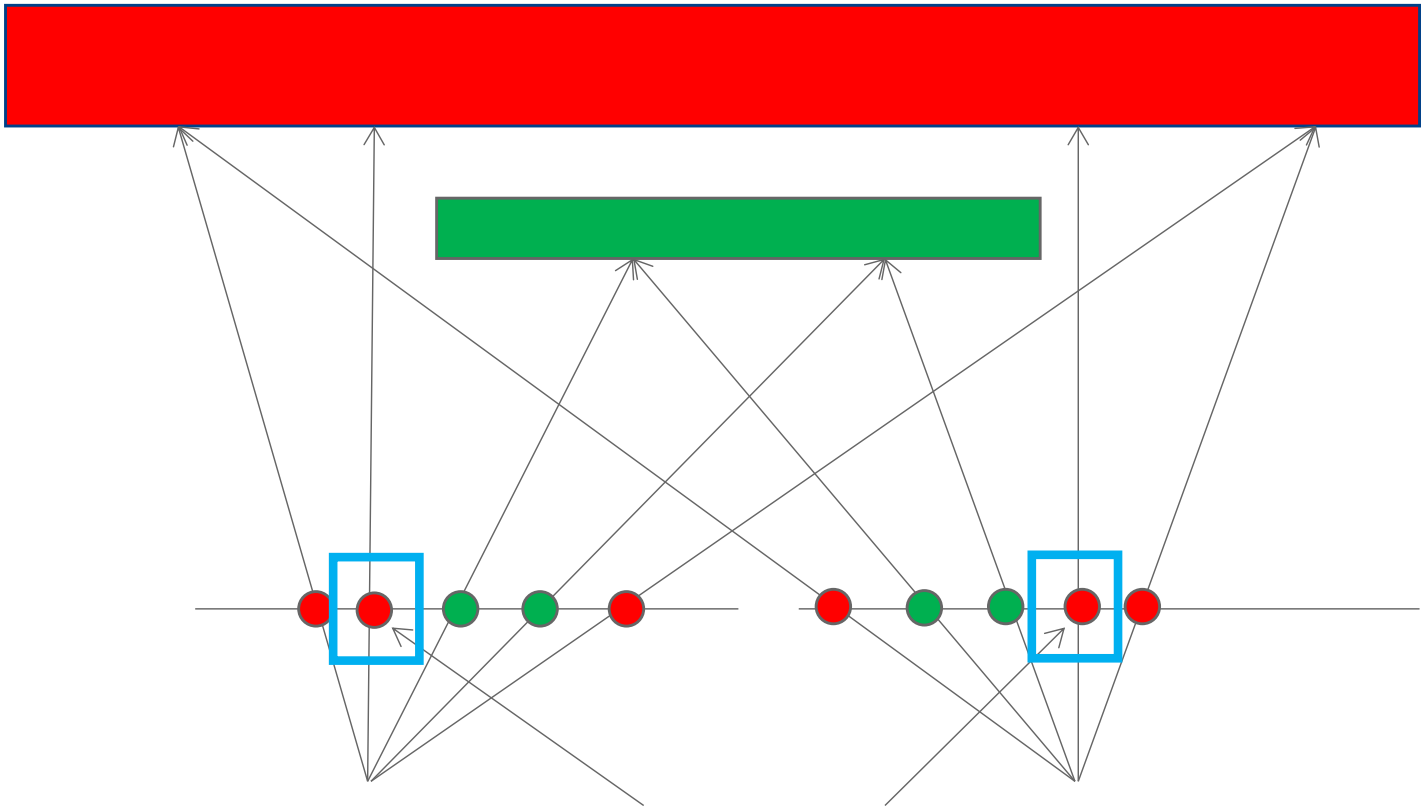


Figure from Gee & Cipolla 1999

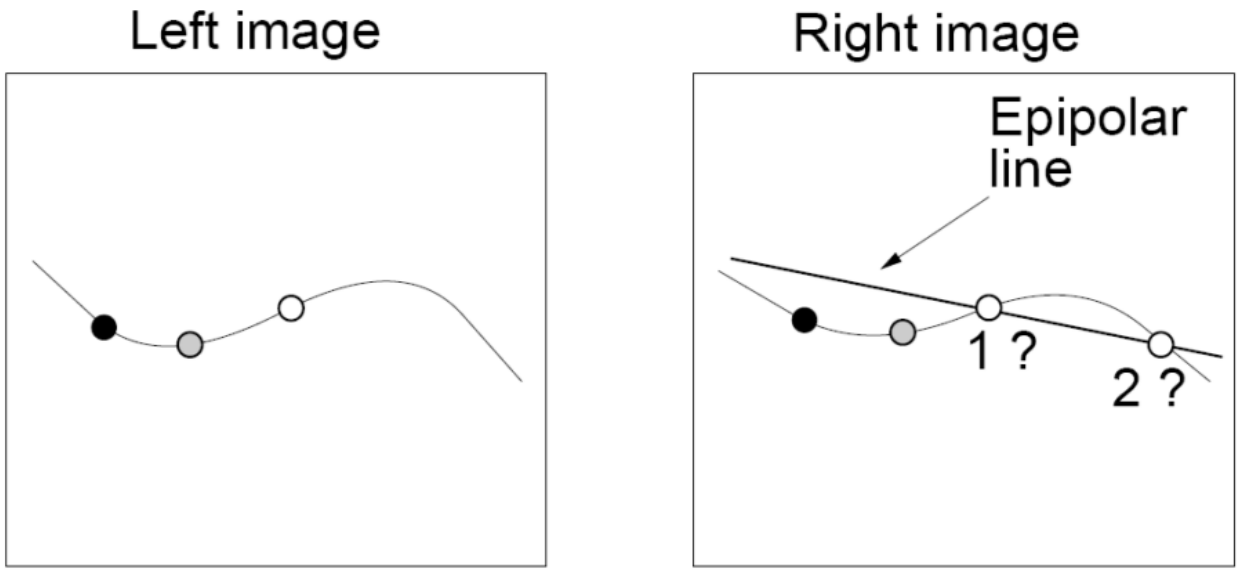
Problem: Occlusion

- Uniqueness says “up to one match” per pixel
- What if there is no match?



Disparity gradient constraint

- Assume piecewise continuous surface, so want disparity estimates to be locally smooth



Given matches ● and ◎, point ○ in the left image must match point 1 in the right image. Point 2 would exceed the disparity gradient limit.

Figure from Gee & Cipolla 1999

Ordering constraint

- Points on **same surface** (opaque object) will be in same order in both views

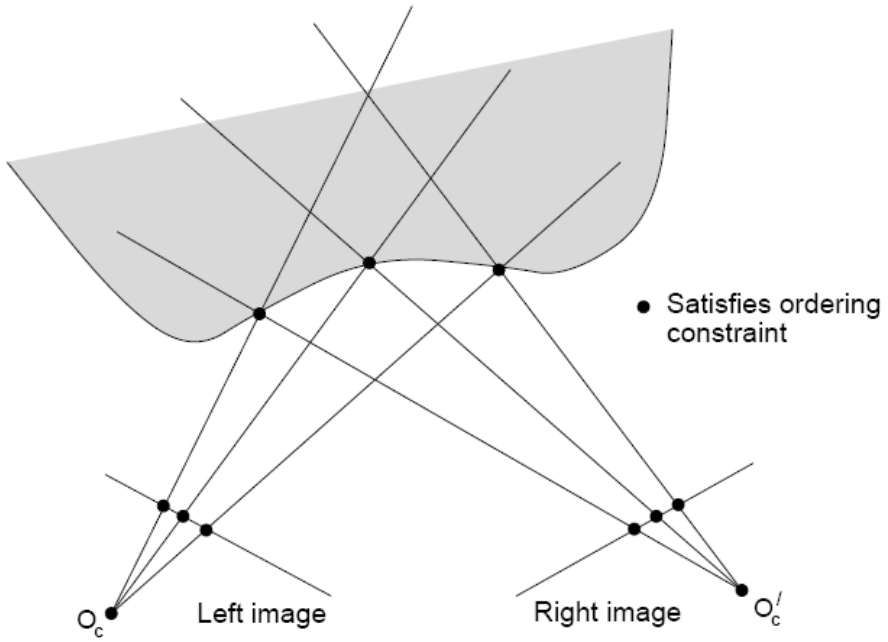
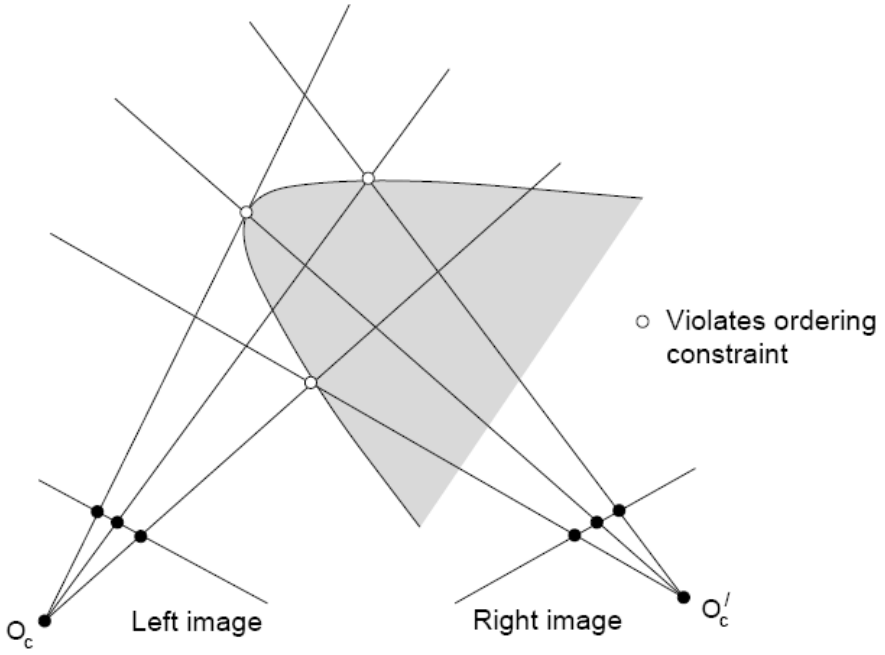
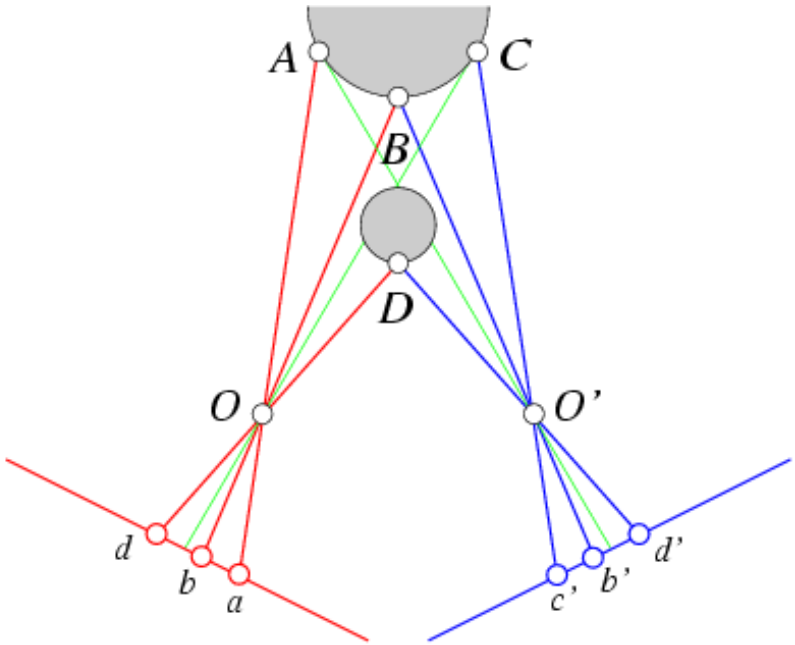


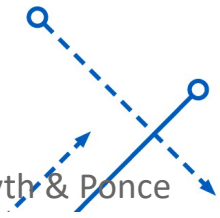
Figure from Gee & Cipolla 1999

Ordering constraint

- Won't always hold, e.g., consider transparent object, or an occluding surface

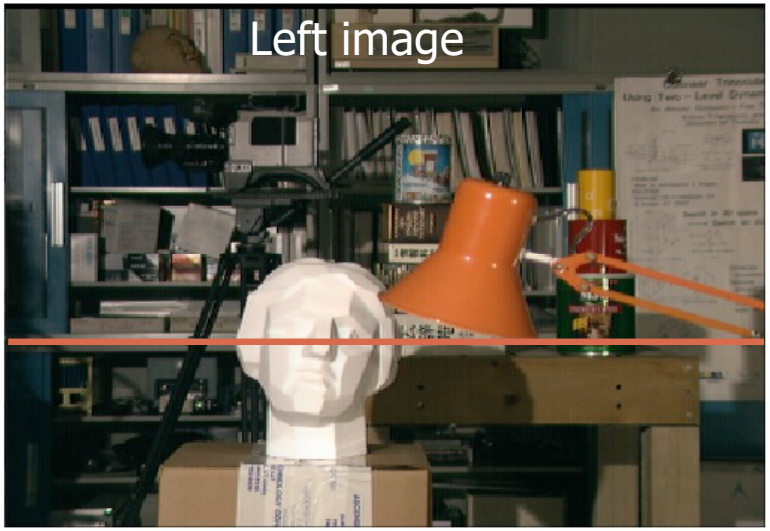


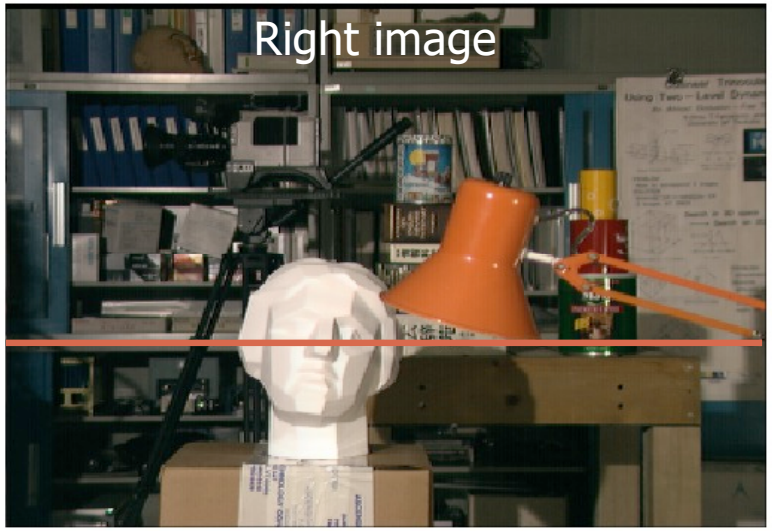


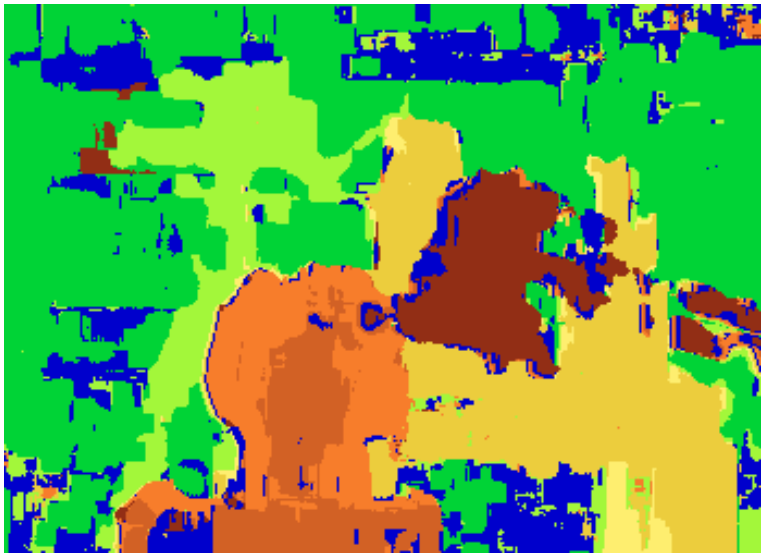


Figures from Forsyth & Ponce

Results with window search





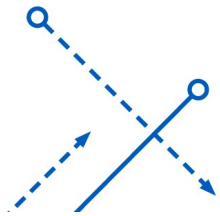


Window-based matching
(best window size)

Ground truth

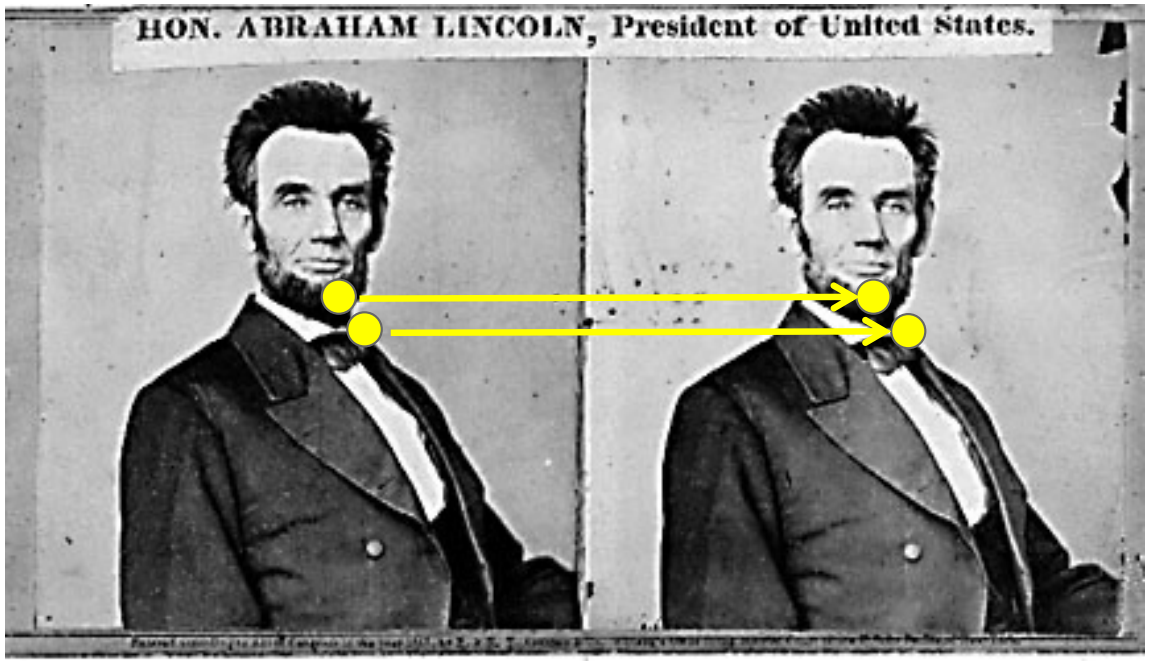
Better solutions

- Beyond individual correspondences estimation
- Optimize correspondence assignments jointly
  - Scanline at a time (DP)
  - Full 2D grid (graph cuts)



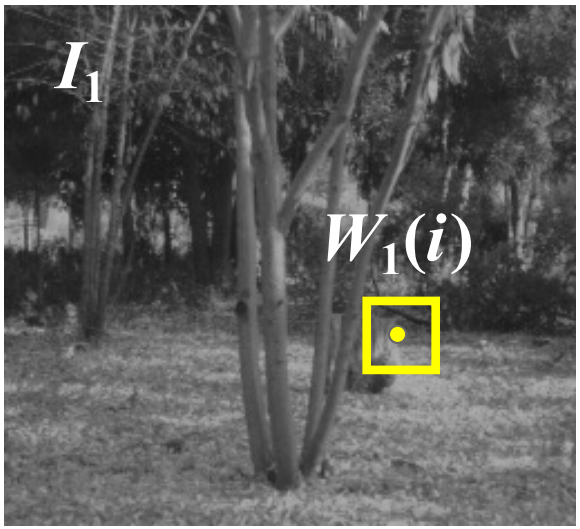
Stereo as energy minimization

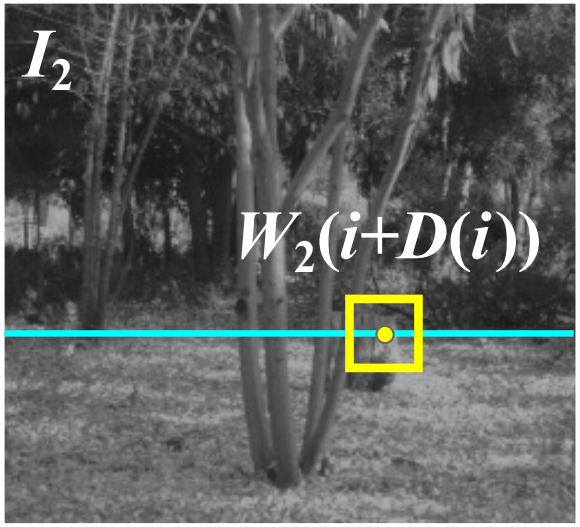
- What defines a good stereo correspondence?
  - Match quality
    - Want each pixel to find a good match in the other image
  - Smoothness
    - Adjacent pixels often move about the same amount.

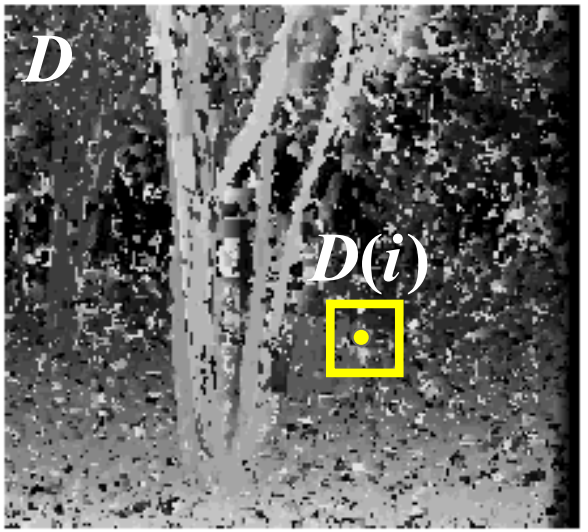


Stereo matching as energy minimization

- Energy functions of this form can be minimized using *graph cuts*







$$E = \alpha E_{\text{data}}(I_1, I_2, D) + \beta E_{\text{smooth}}(D)$$

$$E_{\text{data}} = \sum_i \left(W_1(i) - W_2(i + D(i))\right)^2$$

$$E_{\text{smooth}} = \sum_{\text{neighbors } i,j} \rho\left(D(i) - D(j)\right)$$

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

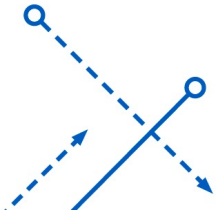
Better results…

Graph cut method

Ground truth

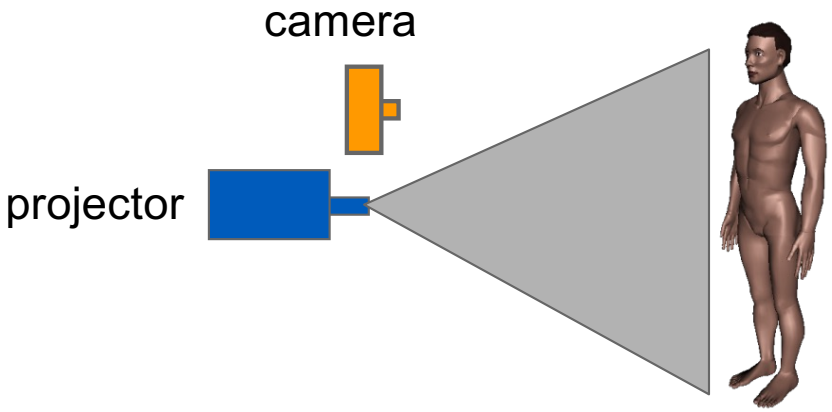
Challenges

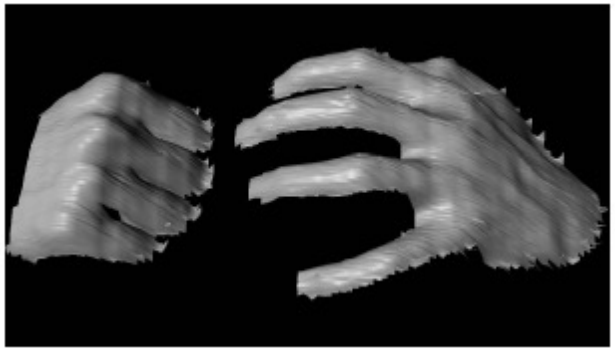
- Low-contrast
  - Textureless image regions
- Occlusions
- Violations of brightness constancy
  - e.g., specular reflections
- Really large baselines
  - Foreshortening and appearance change
- Camera calibration errors



Active stereo with structured light

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





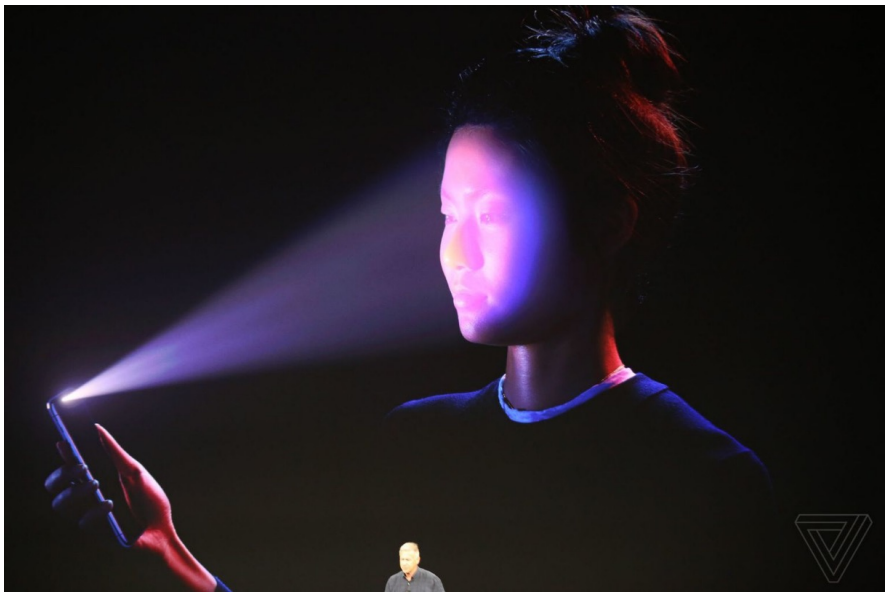
L. Zhang, B. Curless, and S. M. Seitz. Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming. *3DPVT* 2002

Kinect: Structured infrared light

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

iPhone X

- IR Emitter
- 30,000 points
- 2D IR snapshot



SAIR

Spatial AI & Robotics Lab

STEREO VISION

Essential / Fundamental Matrix

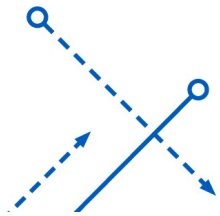
Coordinates in 2-D (Recap)

- Cartesian / homogeneous coordinates of point p

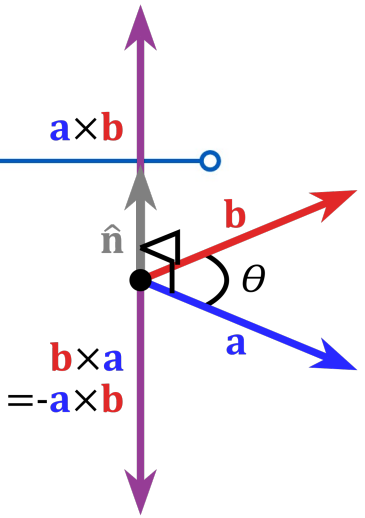
$$\boldsymbol{p} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \qquad \begin{aligned} \boldsymbol{x} &= \boldsymbol{u} / \boldsymbol{w} \\ \boldsymbol{y} &= \boldsymbol{v} / \boldsymbol{w} \end{aligned}$$

- Homogeneous coordinate vector are equivalent if they are proportional to each other

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{u}' \\ \boldsymbol{v}' \\ \boldsymbol{w}' \end{bmatrix} \Leftrightarrow \quad \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{w} \end{bmatrix} \equiv \lambda \begin{bmatrix} \boldsymbol{u}' \\ \boldsymbol{v}' \\ \boldsymbol{w}' \end{bmatrix} \quad \lambda \neq 0$$



Cross product (Recap)



- Cross product of two 3D vector

$$\mathrm{a} \times b = \hat{n}\,|a|\,|b| \sin\theta$$

$$\begin{aligned}\mathbf{a}\times\mathbf{b} &= (a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k})\times(b_1\mathbf{i}+b_2\mathbf{j}+b_3\mathbf{k})\\ &= a_1b_1(\mathbf{i}\times\mathbf{i})+a_1b_2(\mathbf{i}\times\mathbf{j})+a_1b_3(\mathbf{i}\times\mathbf{k})+\\ &\quad a_2b_1(\mathbf{j}\times\mathbf{i})+a_2b_2(\mathbf{j}\times\mathbf{j})+a_2b_3(\mathbf{j}\times\mathbf{k})+\\ &\quad a_3b_1(\mathbf{k}\times\mathbf{i})+a_3b_2(\mathbf{k}\times\mathbf{j})+a_3b_3(\mathbf{k}\times\mathbf{k})\end{aligned}$$

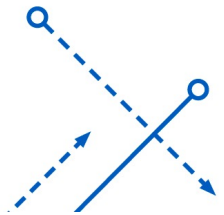
$$\mathbf{a}\times\mathbf{b} = \begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{vmatrix}$$

- Skew-symmetric Matrix

$$\mathbf{a}\times\mathbf{b} = [\mathbf{a}]_\times\mathbf{b} = \begin{bmatrix}0 & -a_3 & a_2\\ a_3 & 0 & -a_1\\ -a_2 & a_1 & 0\end{bmatrix}\begin{bmatrix}b_1\\ b_2\\ b_3\end{bmatrix} \qquad [\mathbf{a}]_\times = \begin{bmatrix}0 & -a_3 & a_2\\ a_3 & 0 & -a_1\\ -a_2 & a_1 & 0\end{bmatrix}$$

- Properties:

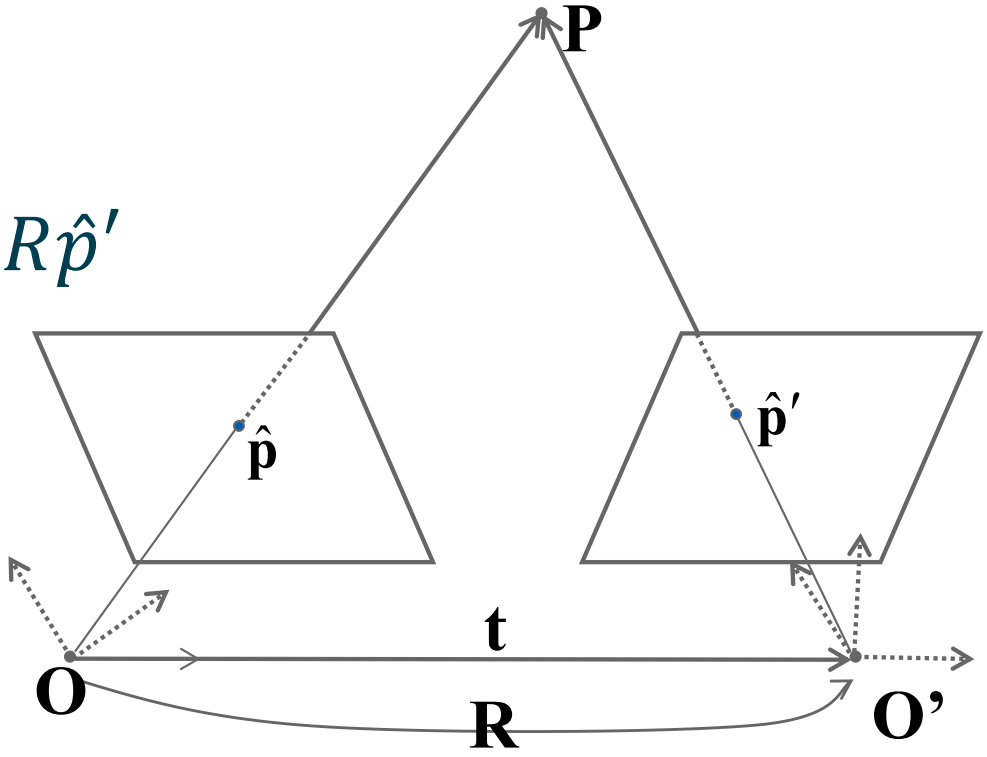
$$\mathrm{a}^{\mathrm{T}}(a\times b) = b^T(a\times b) = 0$$



SAIR
Spatial AI & Robotics Lab

Essential Matrix

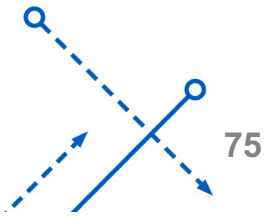
- Due to cross product properties
  - $\overrightarrow{OP} \cdot \left(\overrightarrow{OO'} \times \overrightarrow{O'P}\right) = 0$
- In normalized coordinates:
  - transform O to align O'
- Then direction of $\hat{p}'$ in O is $R\hat{p}'$
  - $\hat{\boldsymbol{p}}^T(\boldsymbol{t} \times R\hat{\boldsymbol{p}}') = 0$
  - $\hat{\boldsymbol{p}}^T([\boldsymbol{t}]_\times R)\hat{\boldsymbol{p}}' = 0$



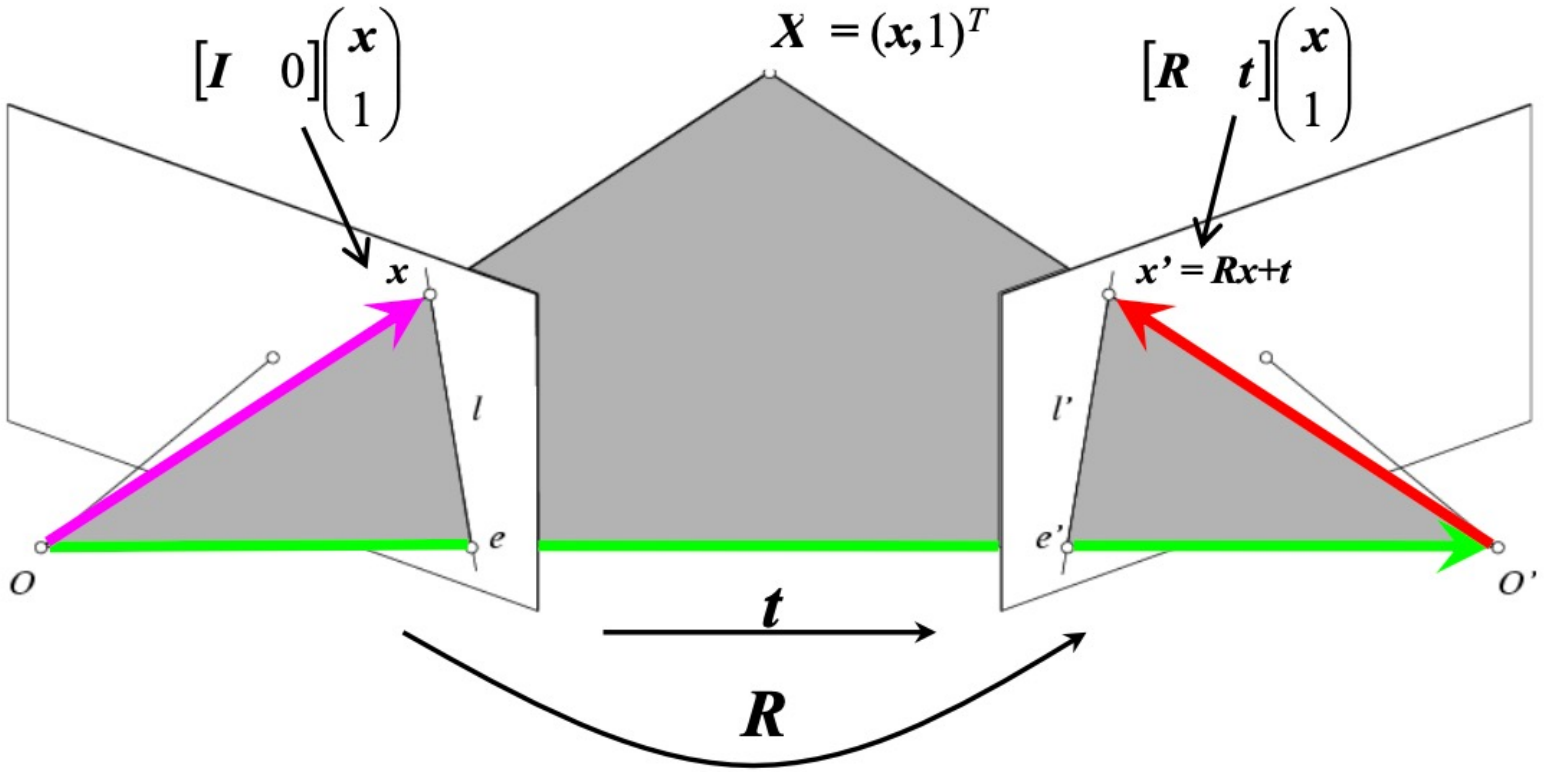
$$\hat{\mathbf{p}}^T \mathbf{E} \hat{\mathbf{p}}' = 0$$

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$$

Essential Matrix



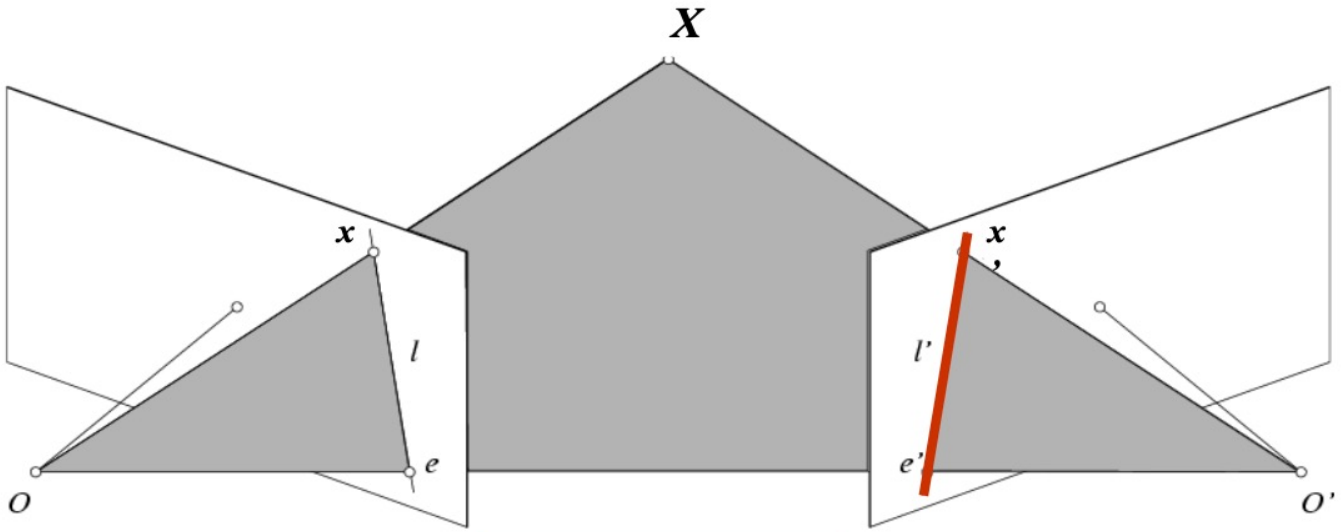
Epipolar constraint: Calibrated case



$$x' \cdot [t \times (Rx)] = 0 \Rightarrow x'^T [t_\times] Rx = 0 \Rightarrow x'^T E x = 0$$

Essential Matrix
(Longuet-Higgins, 1981)

Epipolar constraint: Calibrated case



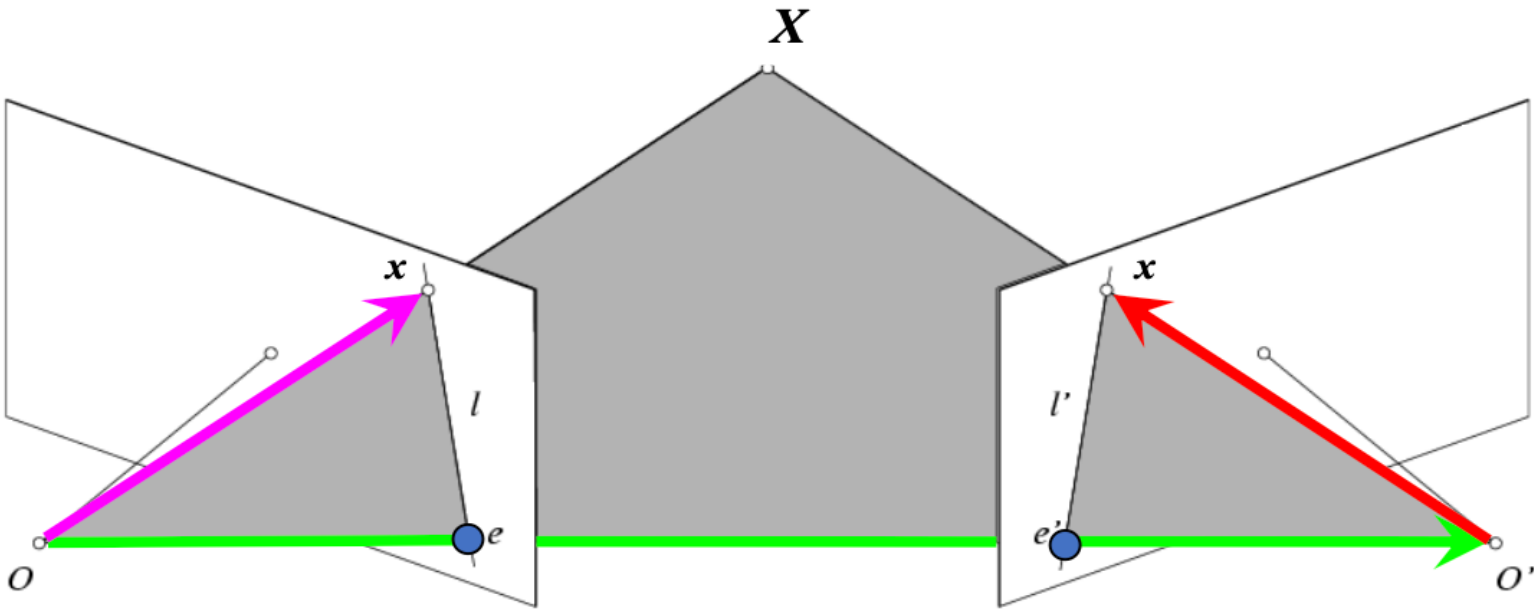
$$\mathbf{x}'^{T} E \mathbf{x} = 0$$

- Ex is the epipolar line associated with x $(l' = Ex)$
- $E^T x'$ is the epipolar line associated with $x'(l = E^T x')$
- $Ee = 0$ and $E^T e' = 0$
- E is **singular** (rank two)
- E has five degrees of freedom

- Recall: a line is given by $ax + by + c = 0$ or

$$\mathbf{l}^T \mathbf{x} = 0 \quad \text{where} \quad \mathbf{l} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

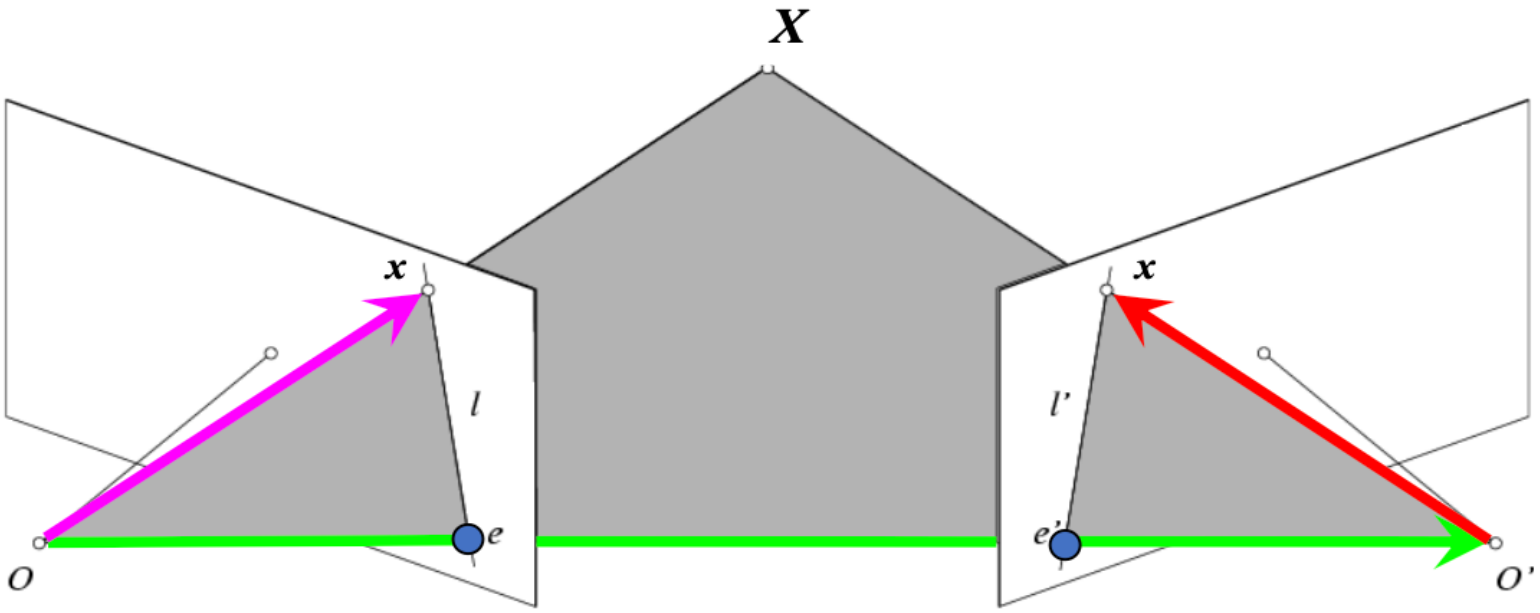
Epipolar constraint: Uncalibrated case



- The calibration matrices K and K' are unknown.
- We can write the epipolar constraint in terms of unknown normalized coordinates:

$$\hat{\boldsymbol{x}}'^{T}\boldsymbol{E}\hat{\boldsymbol{x}}=0 \qquad \hat{\boldsymbol{x}}=\boldsymbol{K}^{-1}\boldsymbol{x}, \quad \hat{\boldsymbol{x}}'=\boldsymbol{K}'^{-1}\boldsymbol{x}'$$

Epipolar constraint: Uncalibrated case



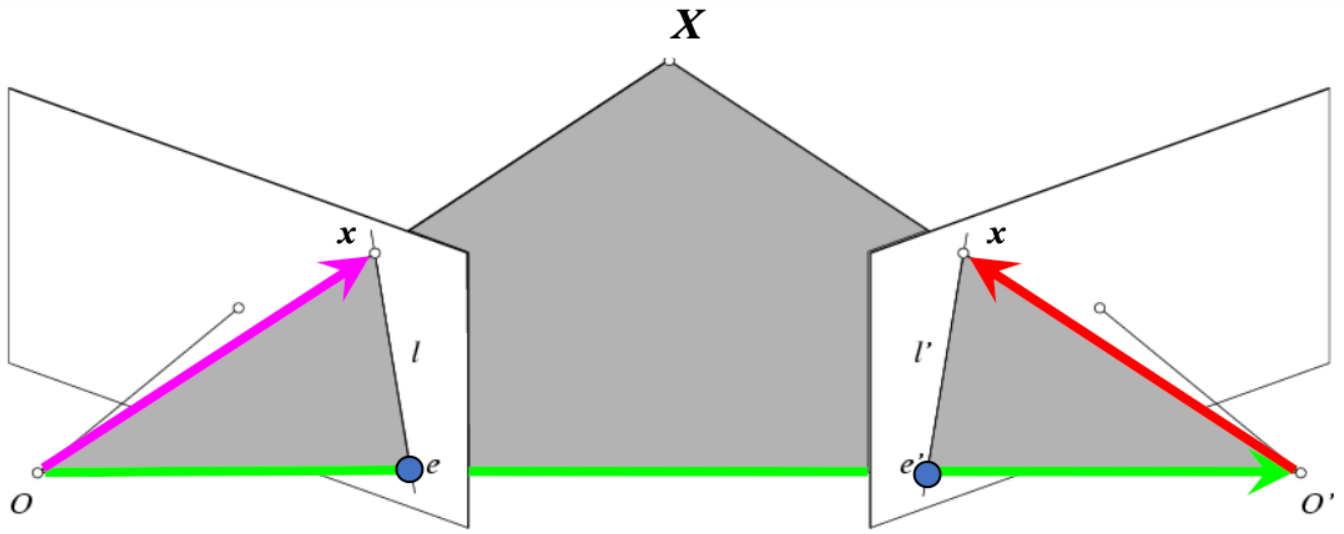
$$\hat{x}'^T E \hat{x} = 0 \quad \Rightarrow \quad x'^T F x = 0 \quad \text{with} \quad F = K'^{-T} E K^{-1}$$

$$\hat{x} = K^{-1} x$$

$$\hat{x}' = K'^{-1} x'$$

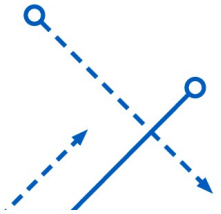
Fundamental Matrix
(Faugeras and Luong, 1992)

Epipolar constraint: Uncalibrated case



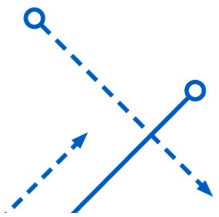
$$\hat{\boldsymbol{x}}'^{T}\boldsymbol{E}\hat{\boldsymbol{x}}=0 \quad \Rightarrow \quad \boldsymbol{x}'^{T}\boldsymbol{F}\boldsymbol{x}=0 \quad \text{with} \quad \boldsymbol{F}=\boldsymbol{K}'^{-T}\boldsymbol{E}\boldsymbol{K}^{-1}$$

- Fx is the epipolar line associated with x $(l' = Fx)$
- $F^T x'$ is the epipolar line associated with x' $(l = F^T x')$
- $Fe = 0$ and $FTe' = 0$
- F is singular (rank two)
- F has seven degrees of freedom



Estimating the Fundamental Matrix

- 8-point algorithm
  - Least squares solution using SVD on equations from 8 pairs of correspondences.
- 7-point algorithm
  - least squares to solve for null space (two vectors) using SVD and 7 pairs of correspondences.
  - Solve for linear combination of null space vectors that satisfies $\det(F) = 0$
- Minimize reprojection error
  - Non-linear least squares
- Note: estimation of F (or E) is degenerate for a planar scene.



8-point algorithm

- Solve a system of homogeneous linear equations
  - a. Write down the system of equations

$$\boldsymbol{x}^T F \boldsymbol{x}' = 0$$

$$uu'f_{11} + uv'f_{12} + uf_{13} + vu'f_{21} + vv'f_{22} + vf_{23} + u'f_{31} + v'f_{32} + f_{33} = 0$$

$$\mathbf{A}\boldsymbol{f} = \begin{bmatrix} u_1u_1' & u_1v_1' & u_1 & v_1u_1' & v_1v_1' & v_1 & u_1' & v_1' & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_nu_v' & u_nv_n' & u_n & v_nu_n' & v_nv_n' & v_n & u_n' & v_n' & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ \vdots \\ f_{33} \end{bmatrix} = \mathbf{0}$$

  - b. Solve f from $Af = 0$ using SVD.

```
[U, S, V] = svd(A);
f = V(:, end);
F = reshape(f, [3 3])';
```

Q & A

- Why do we need 4 points for homography but 7/8 points for fundamental matrix calculation?
  - In the case of fundamental matrix, each point relates to only one constraint, while in homograph, each point is related to two constraints.
- Why can 7 points solve fundamental matrix?
  - In fact, the fundamental matrix only has 7 degrees of freedom. In this case, the rank-2 constraint must be enforced during the computations.