Stereo and Shape-from-Motion Analysis

2D back to 3D

2D to 3D Inference

- Observation
  - Objects are mostly 3D
  - Images are 2D arrays of intensity, color values, etc.
  - 3D depth information is *not explicitly* encoded in *video* images (it is explicitly recorded in *range* images)

2D to 3D Inference (cont.)

- However, 2D analysis implicitly uses 3D info
  - 3D structures are generally not random
    - coherency in motion
  - 3D surfaces of uniform color and reflectivity
    - homogeneous regions in images
  - Man-made objects are of regular shapes and boundaries
    - straight lines and smooth curves in images
- Explicit 3D information can be recovered by examining 2D shape cues
  - disparities in stereo
  - shading change due to orientation
  - texture gradient due to view point change etc.
- Images as “windows” into the 3D world

Shape Inference Techniques

	Passive	Active
Monocular	shape-from-shading, texture, etc.	time-of-flight
Binocular	stereo	laser ranging, structure lighting
Multiple frames	shape-from-motion (SfM, SLAM)	computer tomography, Kinnet

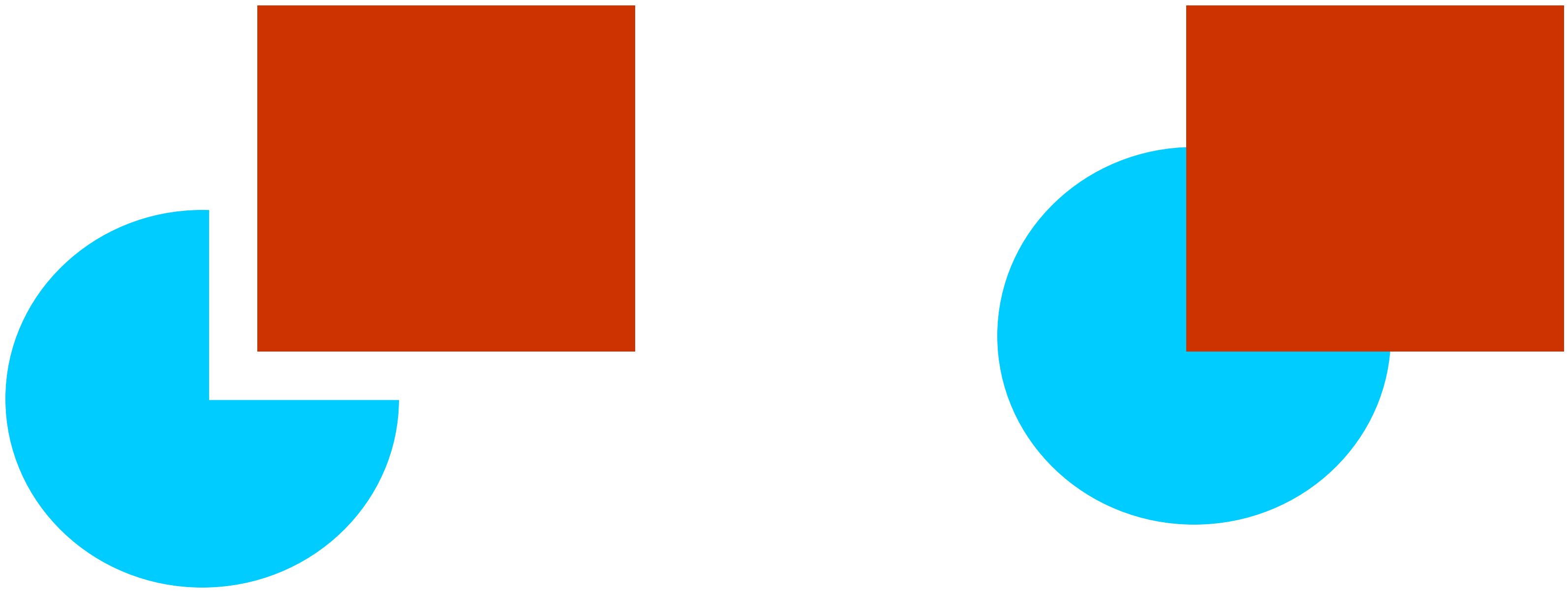
Monocular cues to depth

- **Absolute depth cues**: (assuming known camera parameters) these cues provide information about the absolute depth between the observer and elements of the scene
- **Relative depth cues**: provide relative information about depth between elements in the scene (this point is twice as far at that point, …)

Relative depth cues

Simple and powerful cue, but hard to make it work in practice…

Interposition / occlusion



Texture Gradient

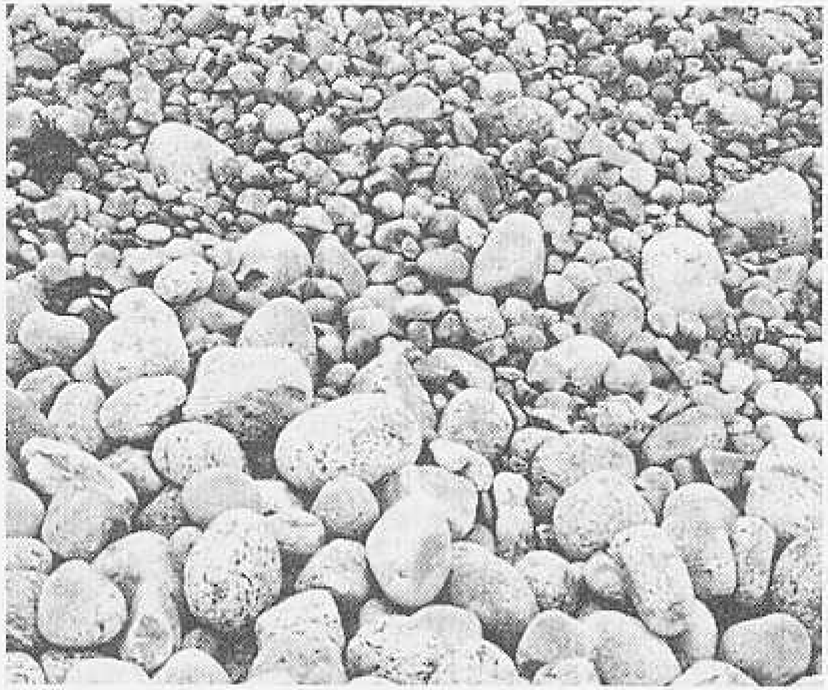


FIGURE 8.27
Texture gradients provide information about depth. (Frank Siteman/Stock, Boston.)

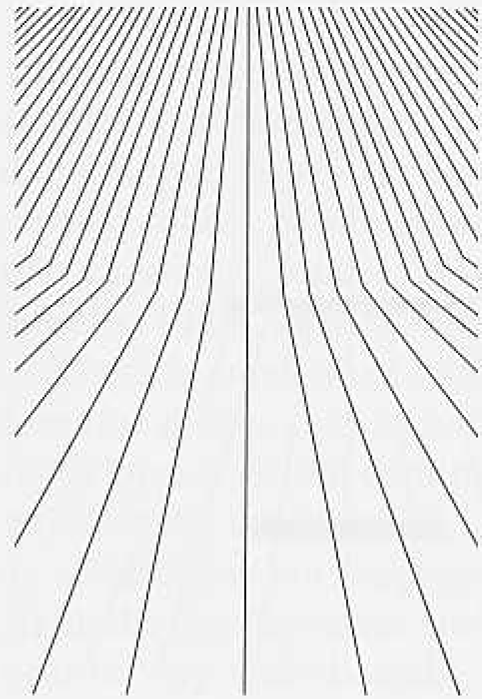
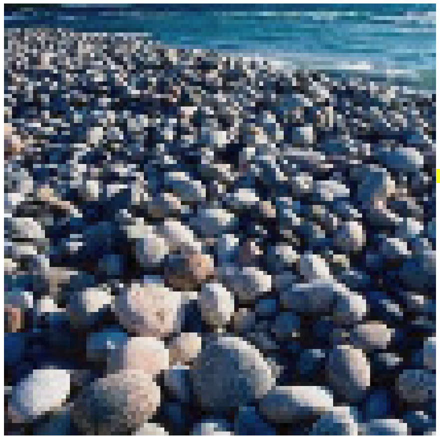
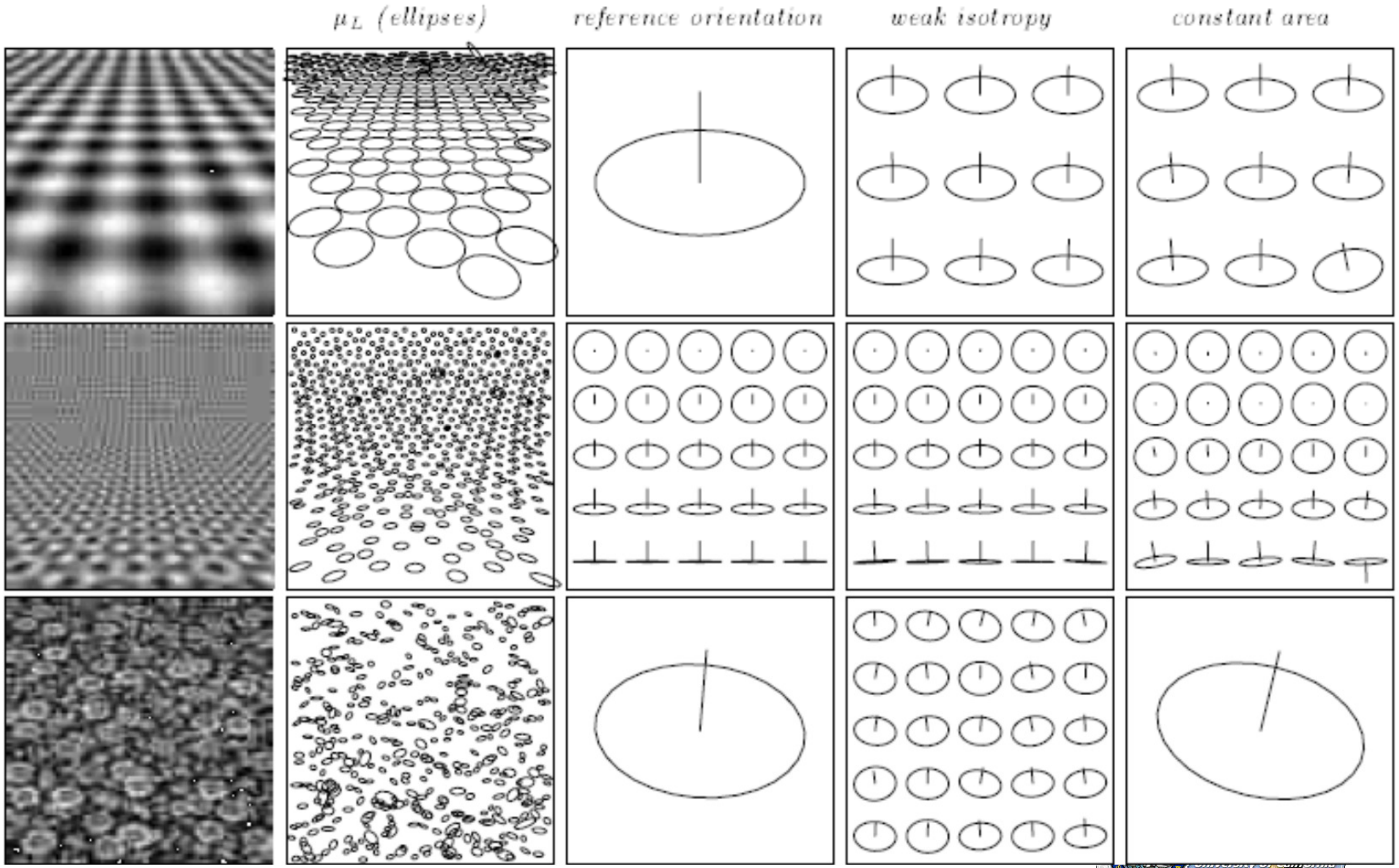


FIGURE 8.28
Texture discontinuity signals the pre
corner.

A Witkin**.** Recovering Surface Shape and Orientation from Texture (1981)



Texture Gradient



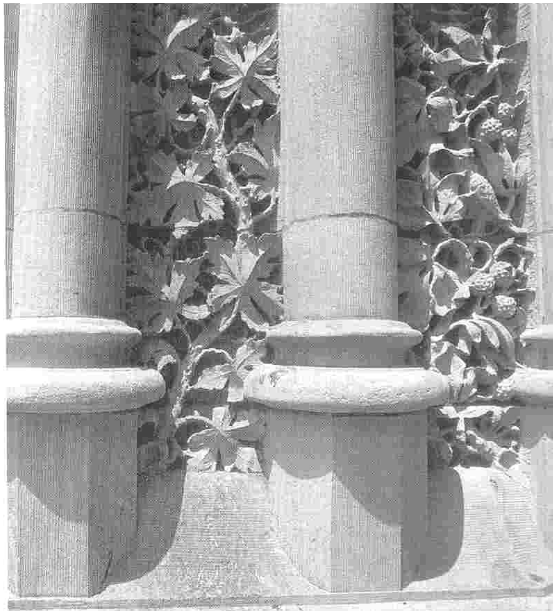
Shape from Texture from a Multi-Scale Perspective. Tony Lindeberg and Jonas Garding. ICCV 93

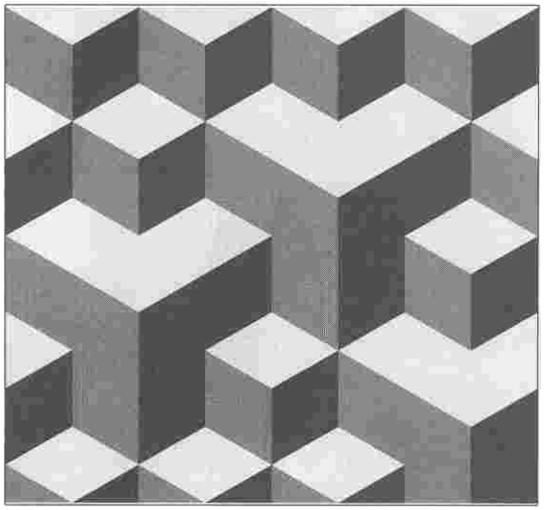
Illumination

- Shading
- Shadows
- Inter-reflections

Shading

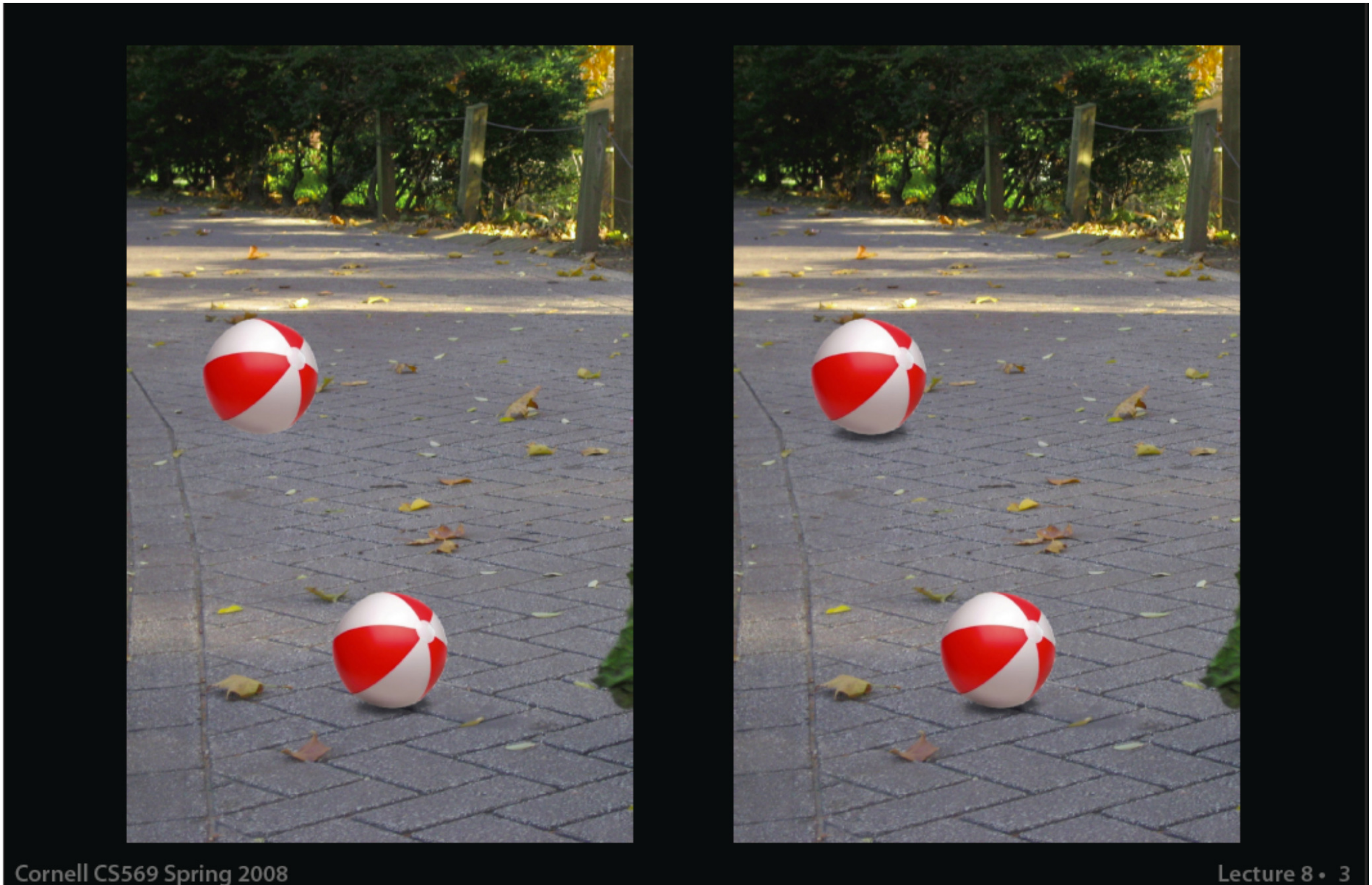
- Based on 3 dimensional modeling of objects in light, shade and shadows.





- Perception of depth through shading alone is always subject to the concave/convex inversion. The pattern shown can be perceived as stairsteps receding towards the top and lighted from above, or as an overhanging structure lighted from below.

Shadows



Slide by Steve Marschner

http://www.cs.cornell.edu/courses/cs569/2008sp/schedule.stm

Linear Perspective

Based on the apparent convergence of parallel lines to common vanishing points with increasing distance from the observer.

(Gibson : “perspective order”)

In Gibson’s term, perspective is a characteristic of the visual field rather than the visual world. It approximates how we see (the retinal image) rather than what we see, the objects in the world.

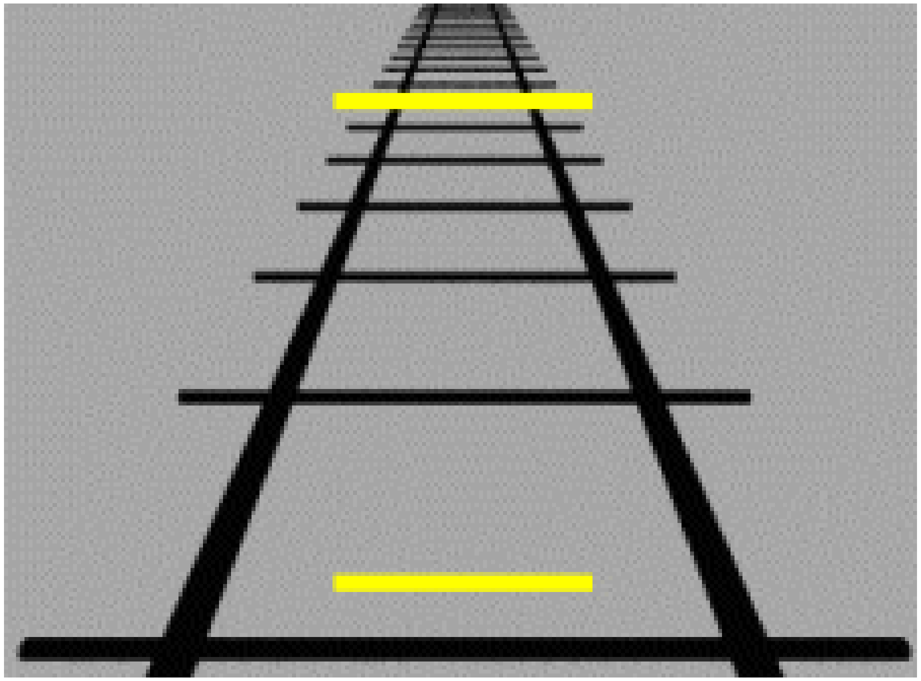
Perspective : a representation that is specific to one individual, in one position in space and one moment in time (a powerful immediacy).

Is perspective a universal fact of the visual retinal image ? Or is perspective something that is learned ?

Simple and powerful cue, and easy to make it work in practice….

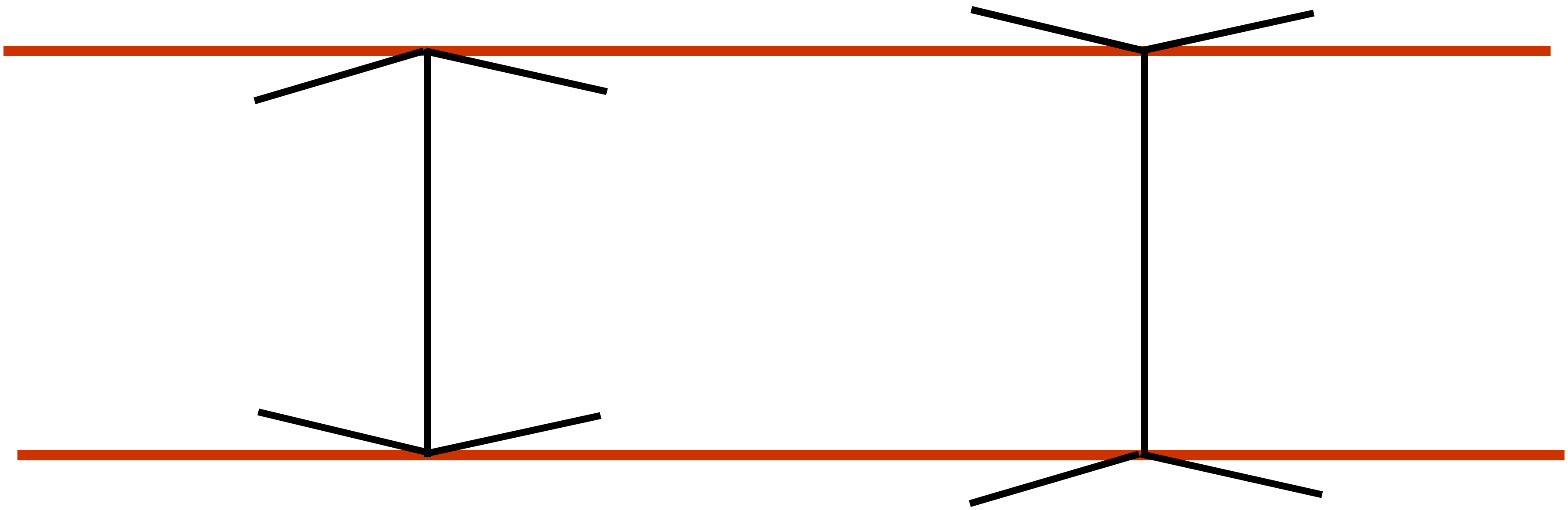
University of California Santa Barbara

Linear Perspective



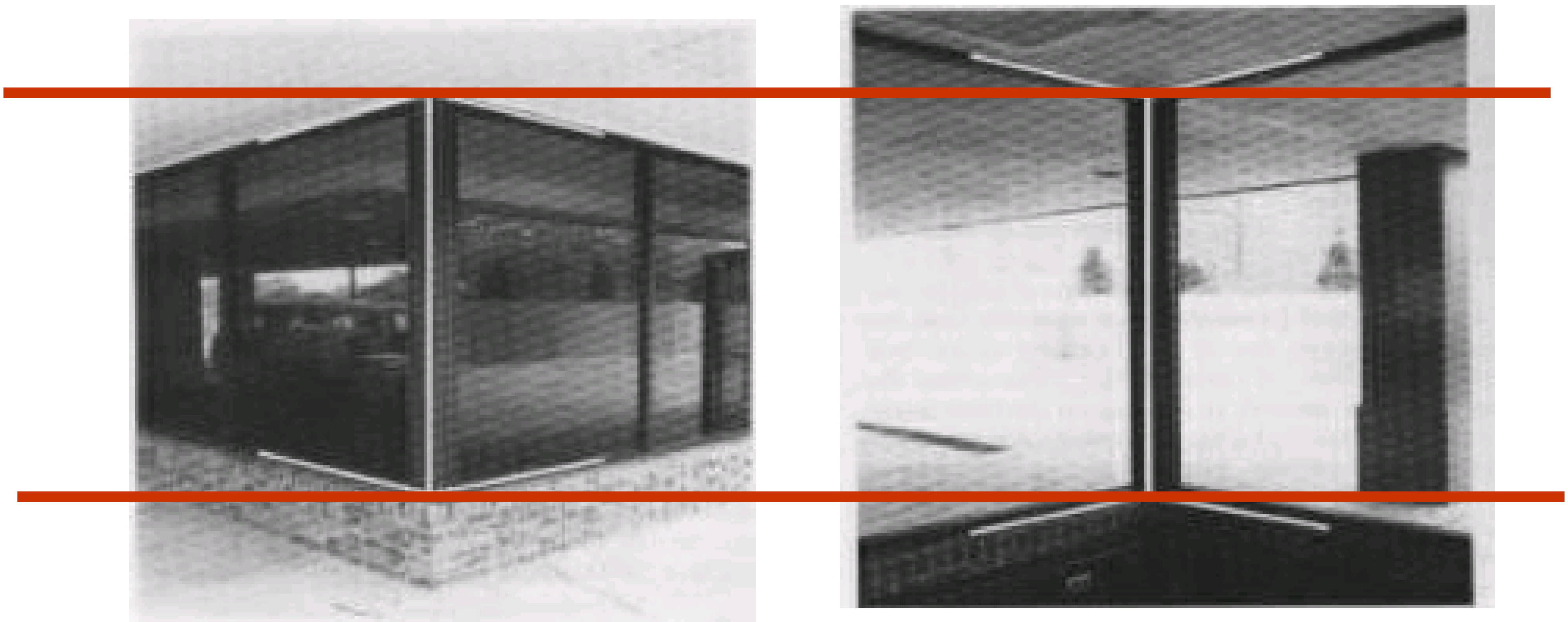
Ponzo's illusion

Linear Perspective



Muller-Lyer
1889

Linear Perspective

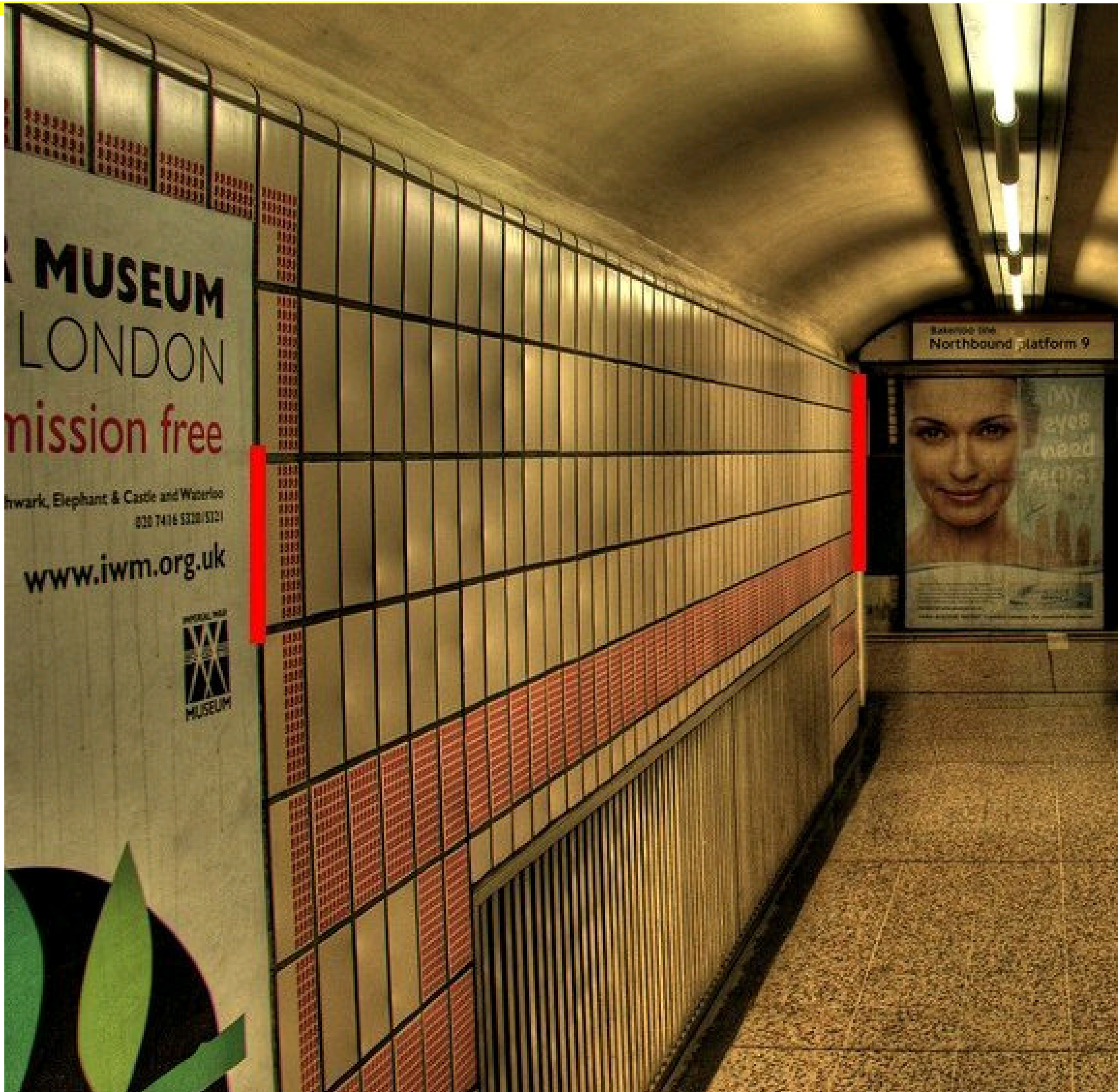


Muller-Lyer
1889

Linear Perspective

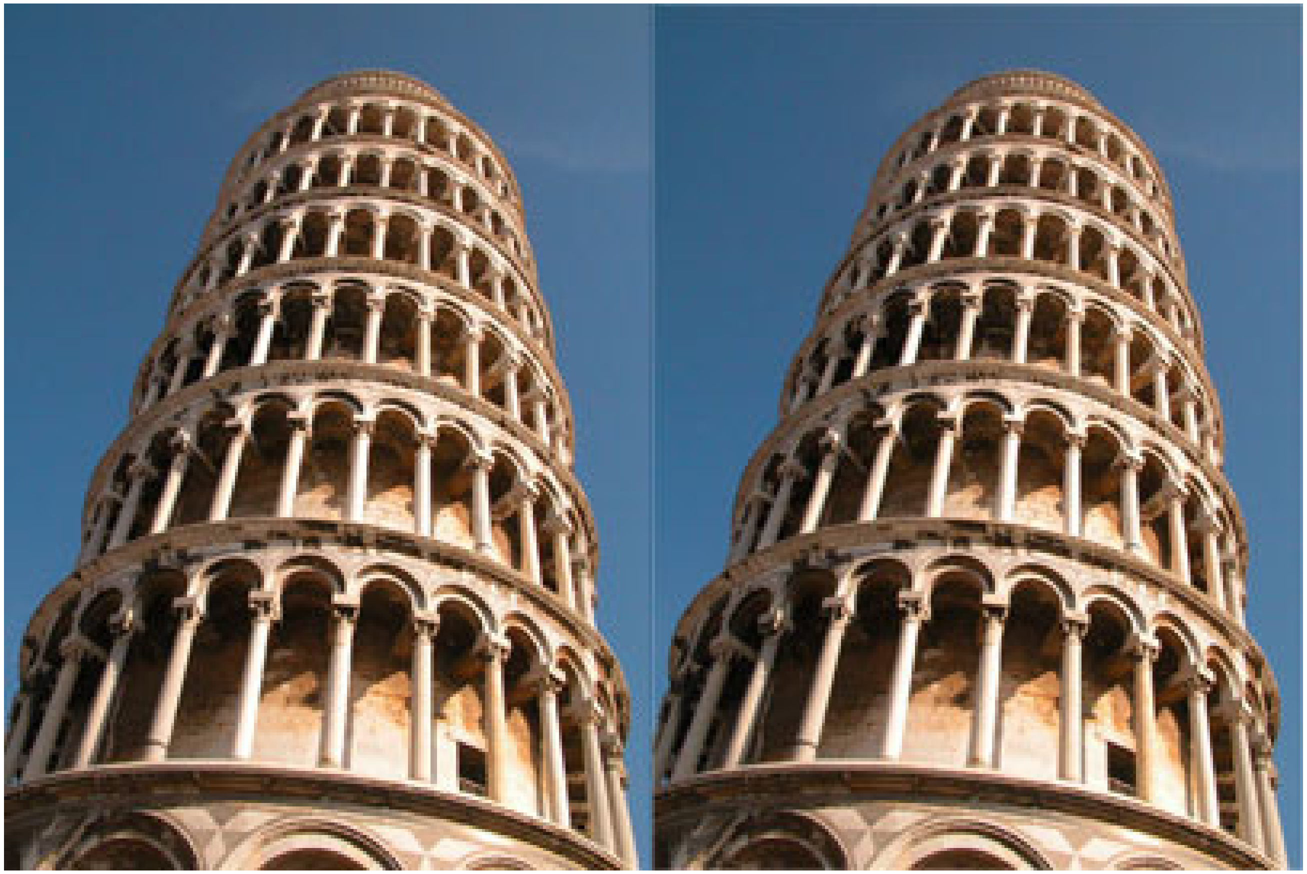
Muller-Lyer
1889

Linear Perspective



(c) 2006 Walt Anthony

3D drives perception of important object attributes

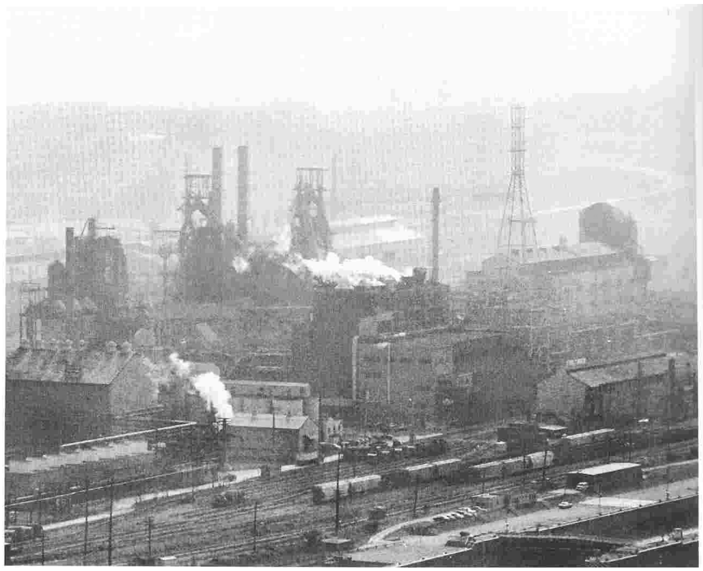


The two Towers of Pisa

Frederick Kingdom, Ali Yoonessi and Elena Gheorghiu of McGill Vision Research unit.

Atmospheric perspective

- Based on the effect of air on the color and visual acuity of objects at various distances from the observer.
- Consequences:
  - Distant objects appear bluer
  - Distant objects have lower contrast.



Atmospheric perspective

http://encarta.msn.com/medias_761571997/Perception_(psychology).html

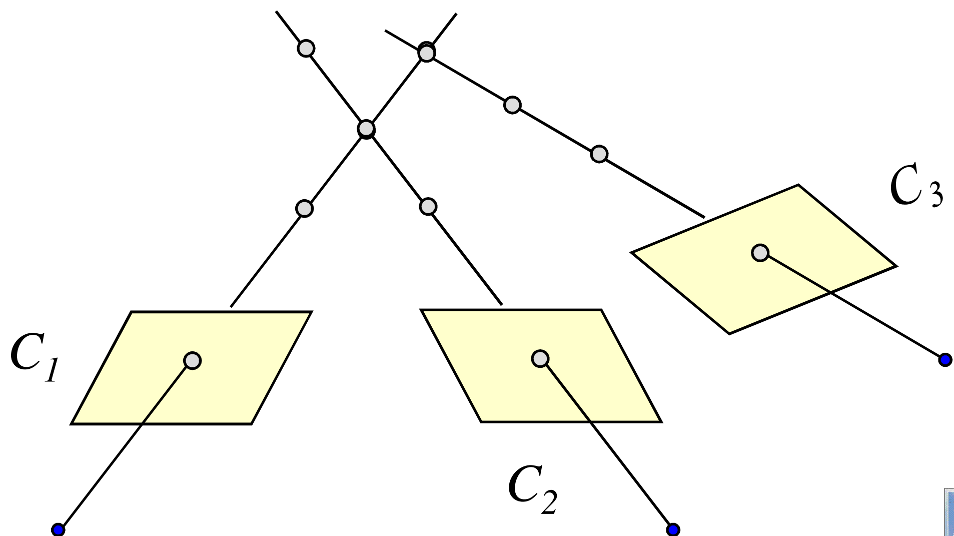
Claude Lorrain (artist)
French, 1600 - 1682
***Landscape with Ruins, Pastoral Figures, and Trees*, 1643/1655**

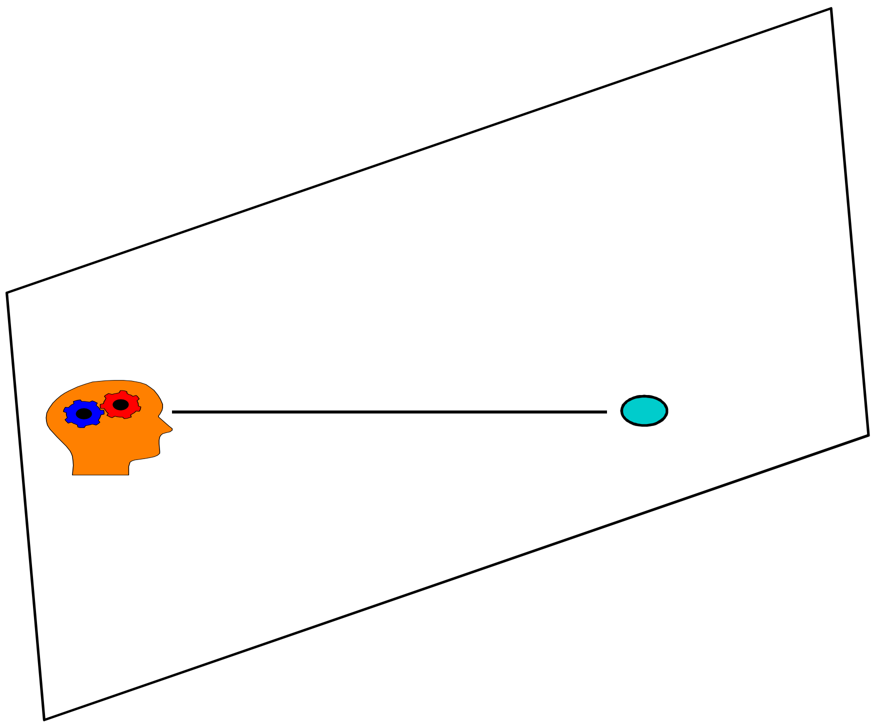
Why multiple views?

- There are many cues to depth and 3D structure besides stereo
  - Oculomotor convergence/divergence, accomodation (changing focus), motion parallax (changing viewpoint)
  - Monocular depth cues (occlusion, perspective, texture gradients, shading, size)
- Multiple views are not always needed – humans can figure out a lot from a single 2D view!

Why multiple views?

- But precise 3D information (distance, depth, shape, curvature, etc.) is difficult or impossible to obtain from a single view
- In order to measure distances, sizes, angles, etc. we need multiple views (and calibrated cameras!)
  - Monocular → binocular → trinocular…





Fixation, convergence

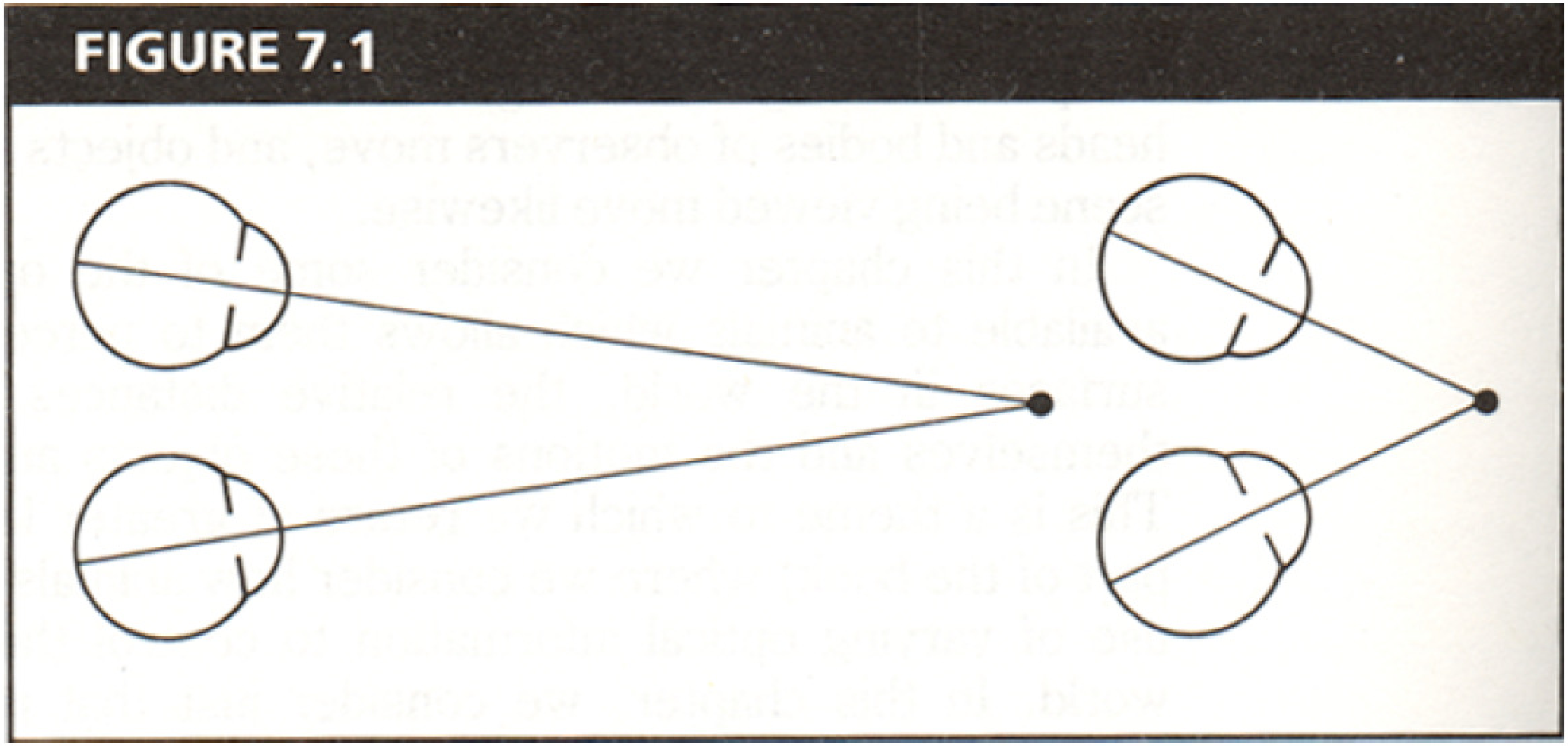
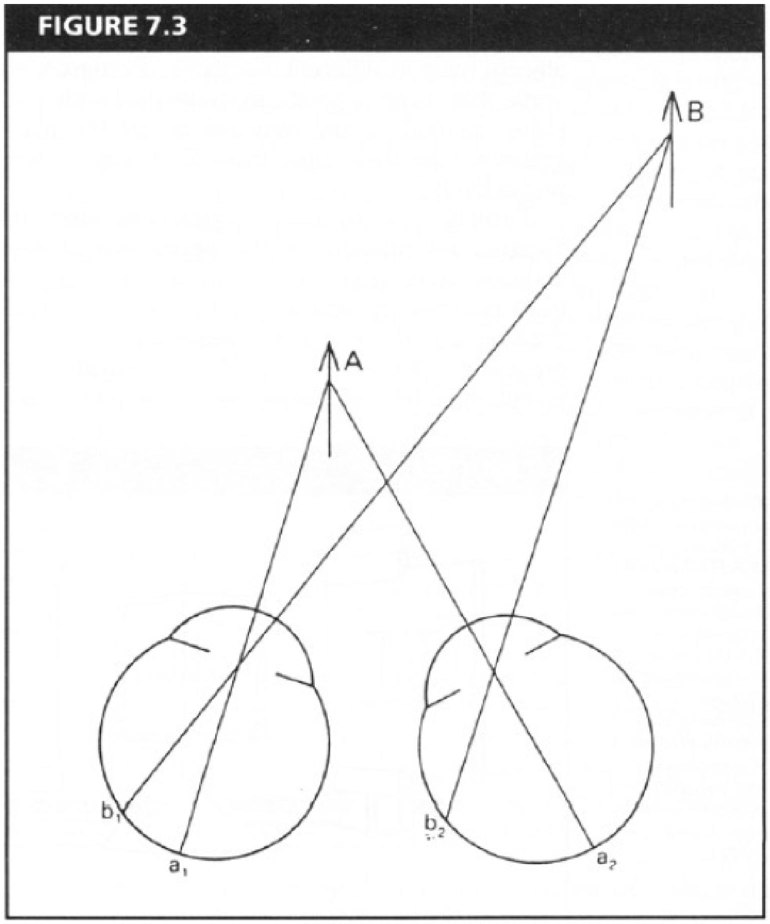


FIGURE 7.1

From Bruce and Green, Visual Perception, Physiology, Psychology and Ecology

Grauman

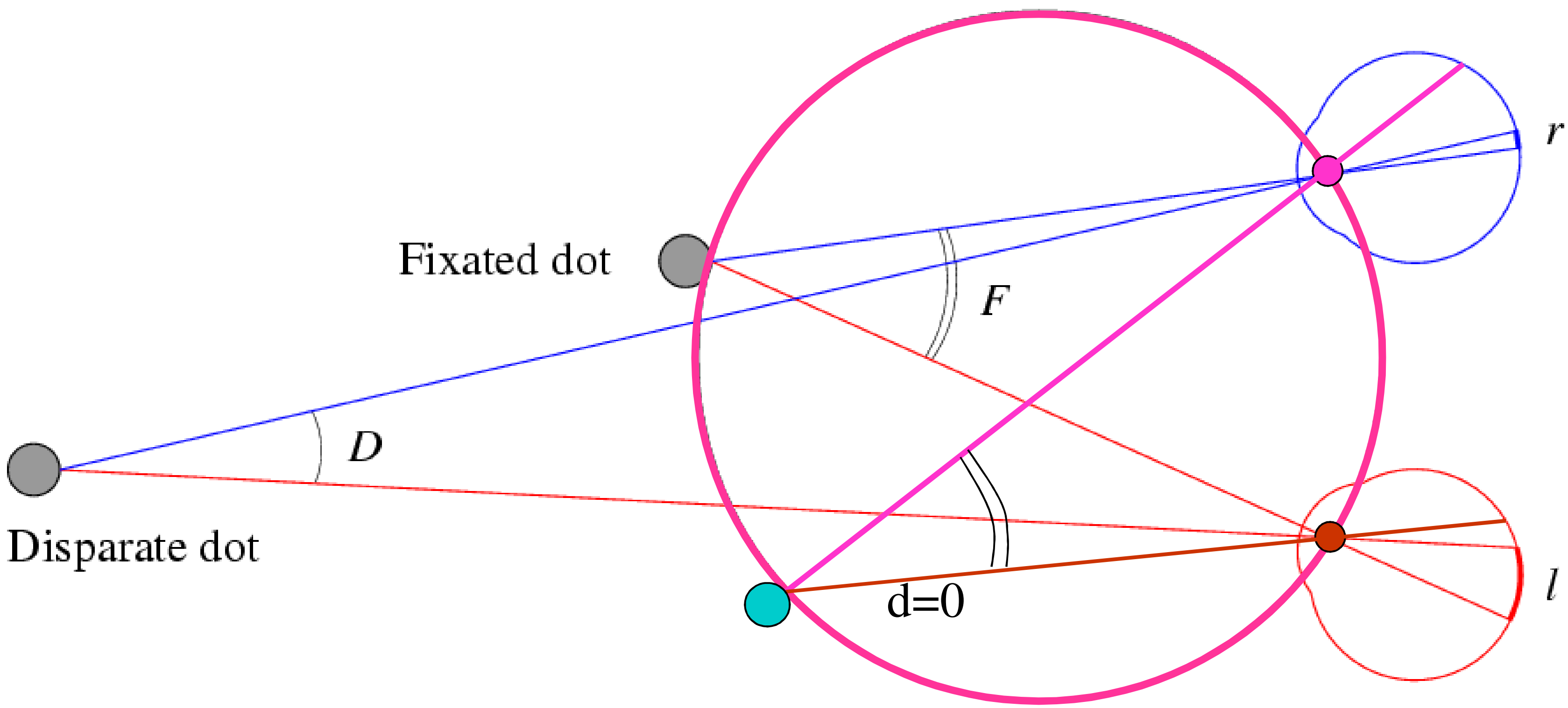
Human stereopsis: disparity



From Bruce and Green, Visual Perception, Physiology, Psychology and Ecology

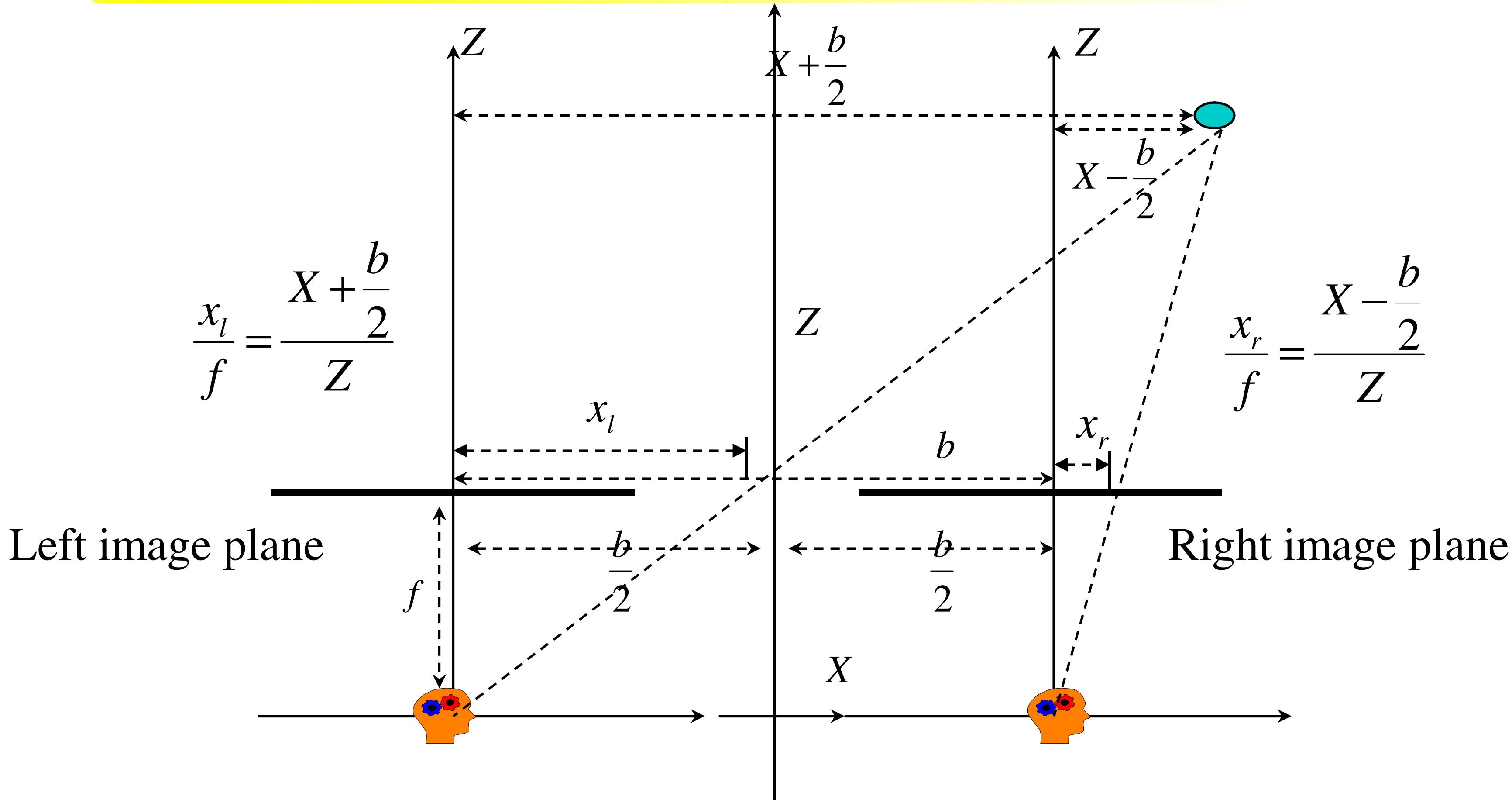
Disparity occurs when eyes fixate on one object; others appear at different visual angles

Human stereopsis: disparity



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

Basic Stereo Configuration



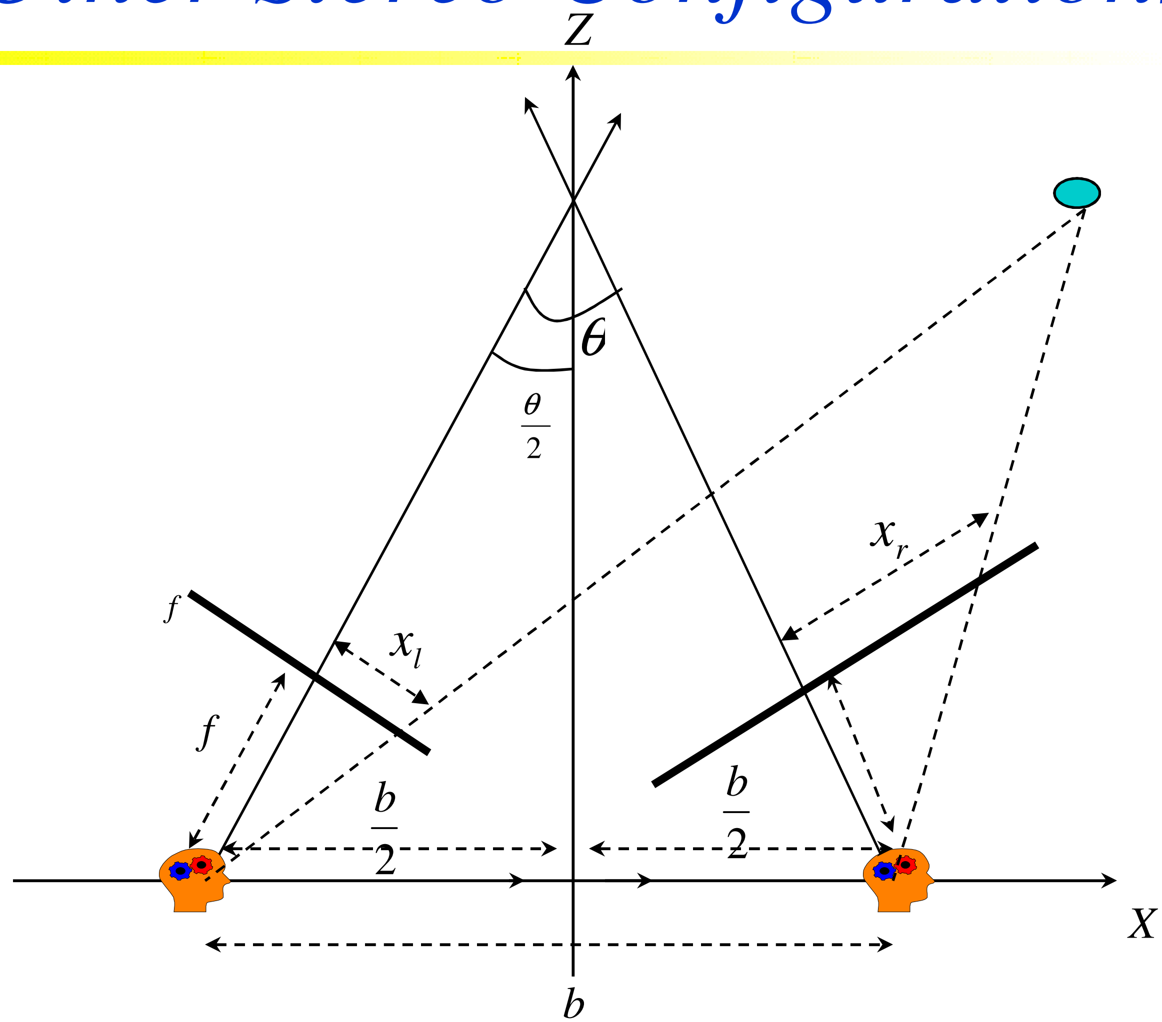
$$\frac{x_l}{f} = \frac{X + \frac{b}{2}}{Z}$$

$$\frac{x_r}{f} = \frac{X - \frac{b}{2}}{Z}$$

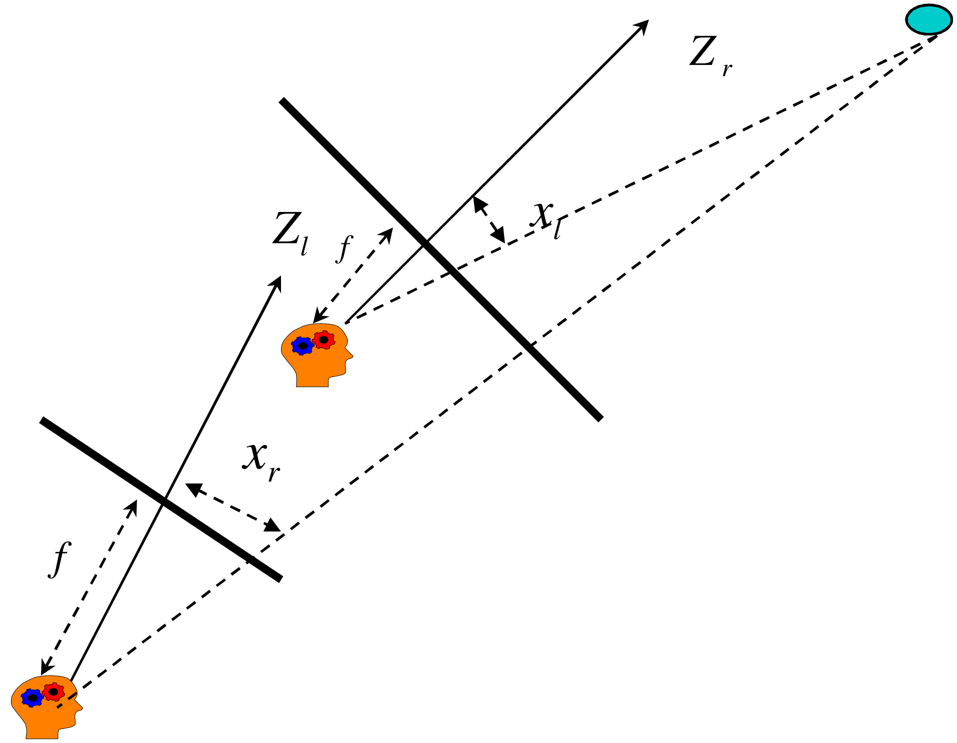
$$\frac{x_l - x_r}{f} = \frac{b}{Z}$$

$$Z = \frac{bf}{(x_l - x_r)}$$

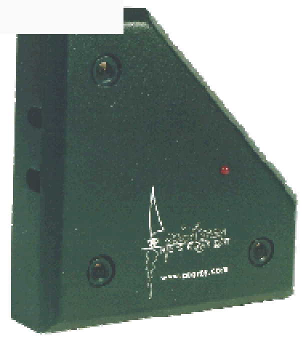
Other Stereo Configurations



Yet Other Stereo Configurations



Stereo camera examp

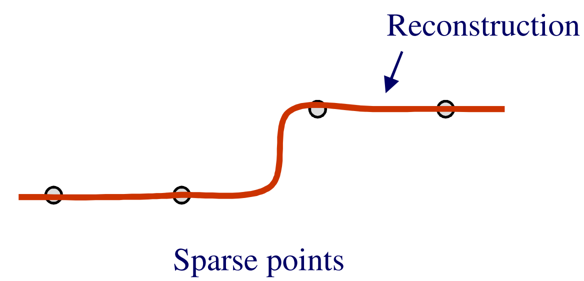


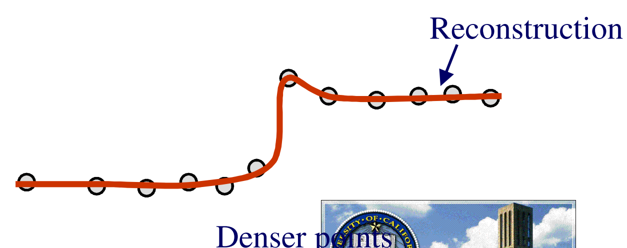
Basic approach to stereo vision

- Find features of interest in *N* image views
  - The “correspondence problem”

- Triangulate from pairs of views
  - A method to measure distance and direction by forming a triangle and using trigonometry

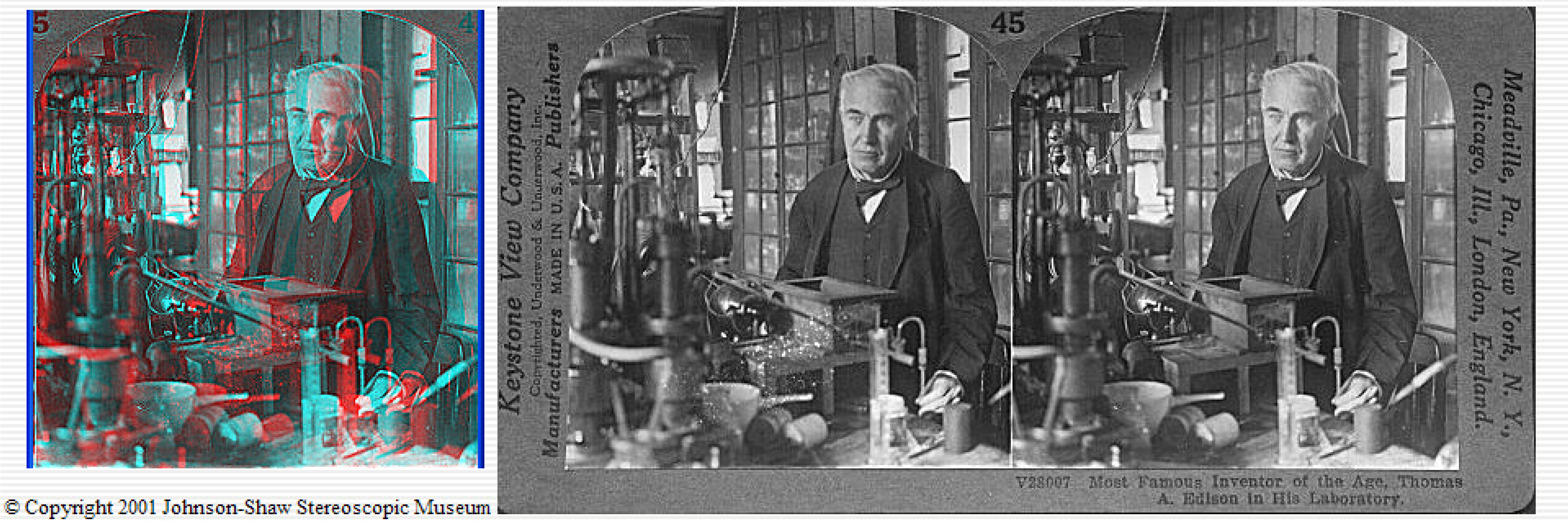
- Reconstruct object/scene depth
  - From dense points
  - From sparse points



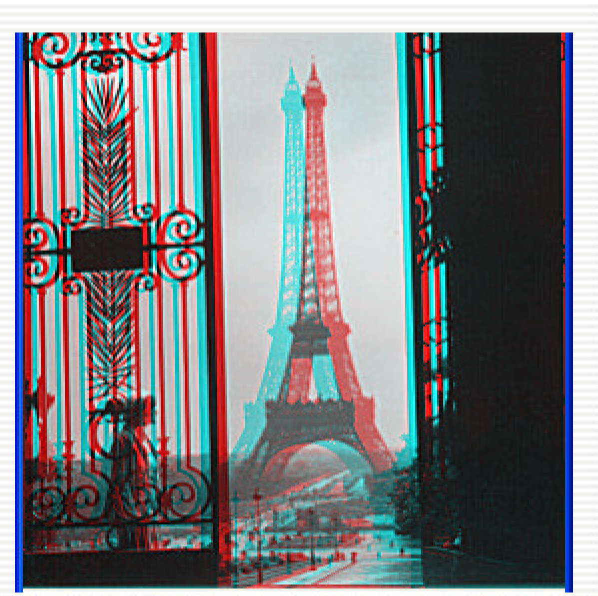


Shape-from-Motion Analysis

- What do we know?
  - Unfortunately, very little beyond feature correspondence (again, inferred, not given)
  - We do NOT know baseline (or how cameras move)
  - UAV flight, mobile robots, hand-held cameras, etc
- This is a MUCH harder problem than stereo
- Stereo can theoretically recover absolute scale while SfM cannot
- We lump them together because the math is the same (stereo does not need to infer camera motion)



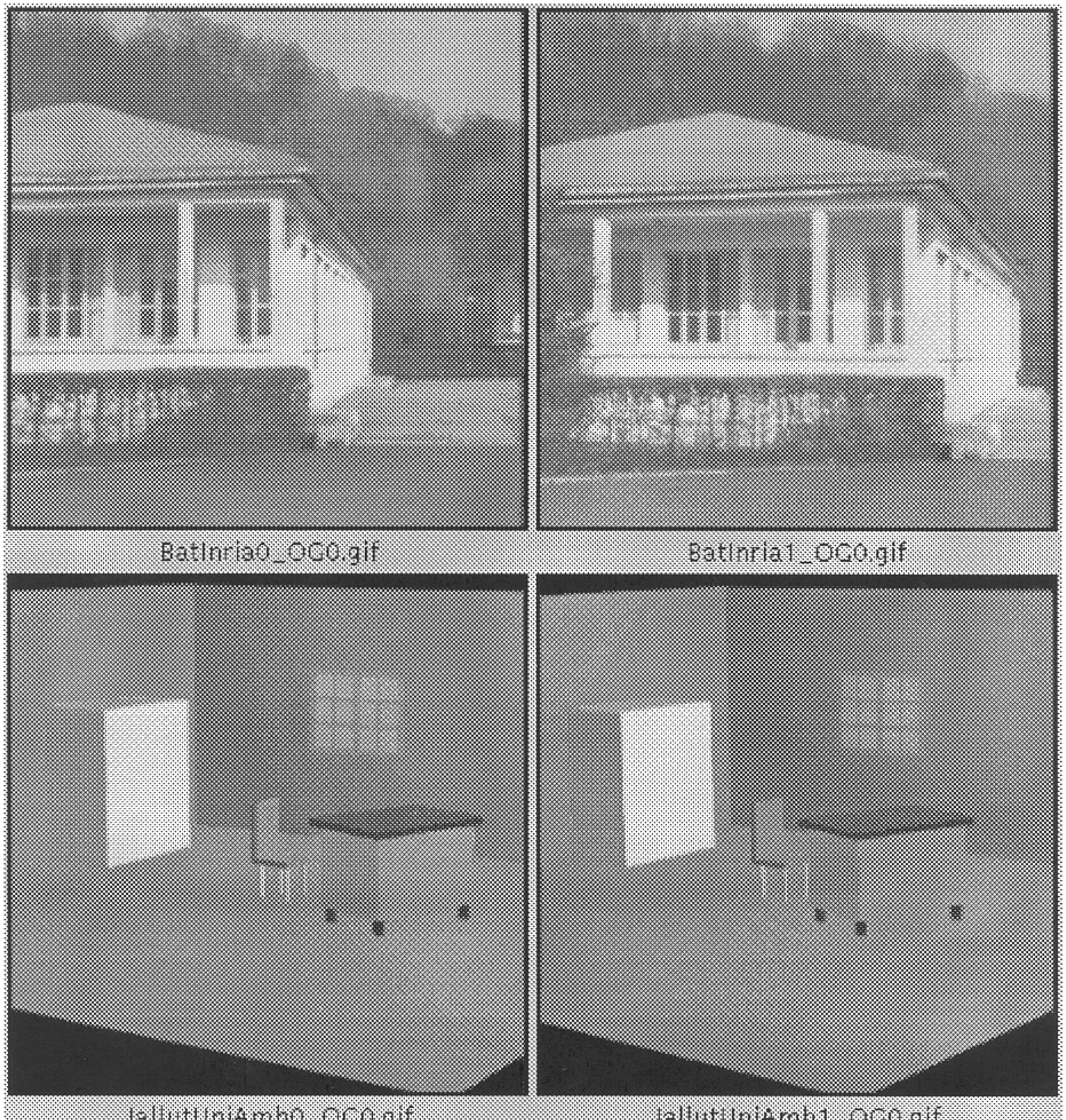
http://www.johnsonshawmuseum.org

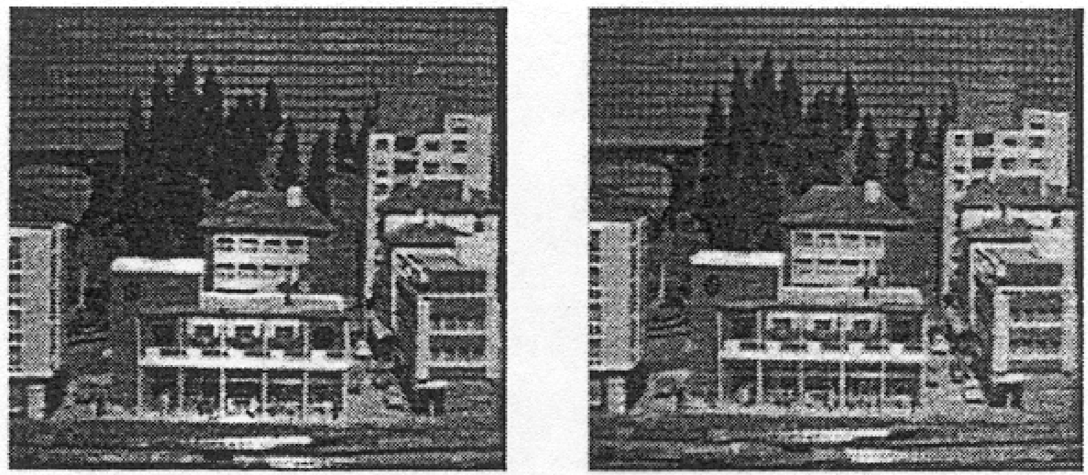


http://www.johnsonshawmuseum.org

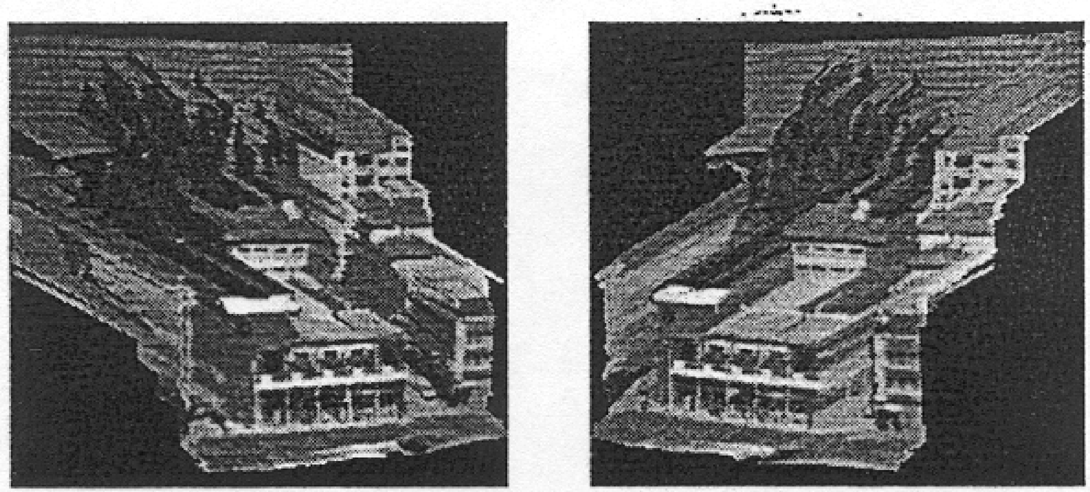
Public Library, Stereoscopic Looking Room, Chicago, by Phillips, 1923

Grauman


BatInria0_OG0.gif
BatInria1_OG0.gif
JallutUniAmb0_OG0.gif
JallutUniAmb1_OG0.gif

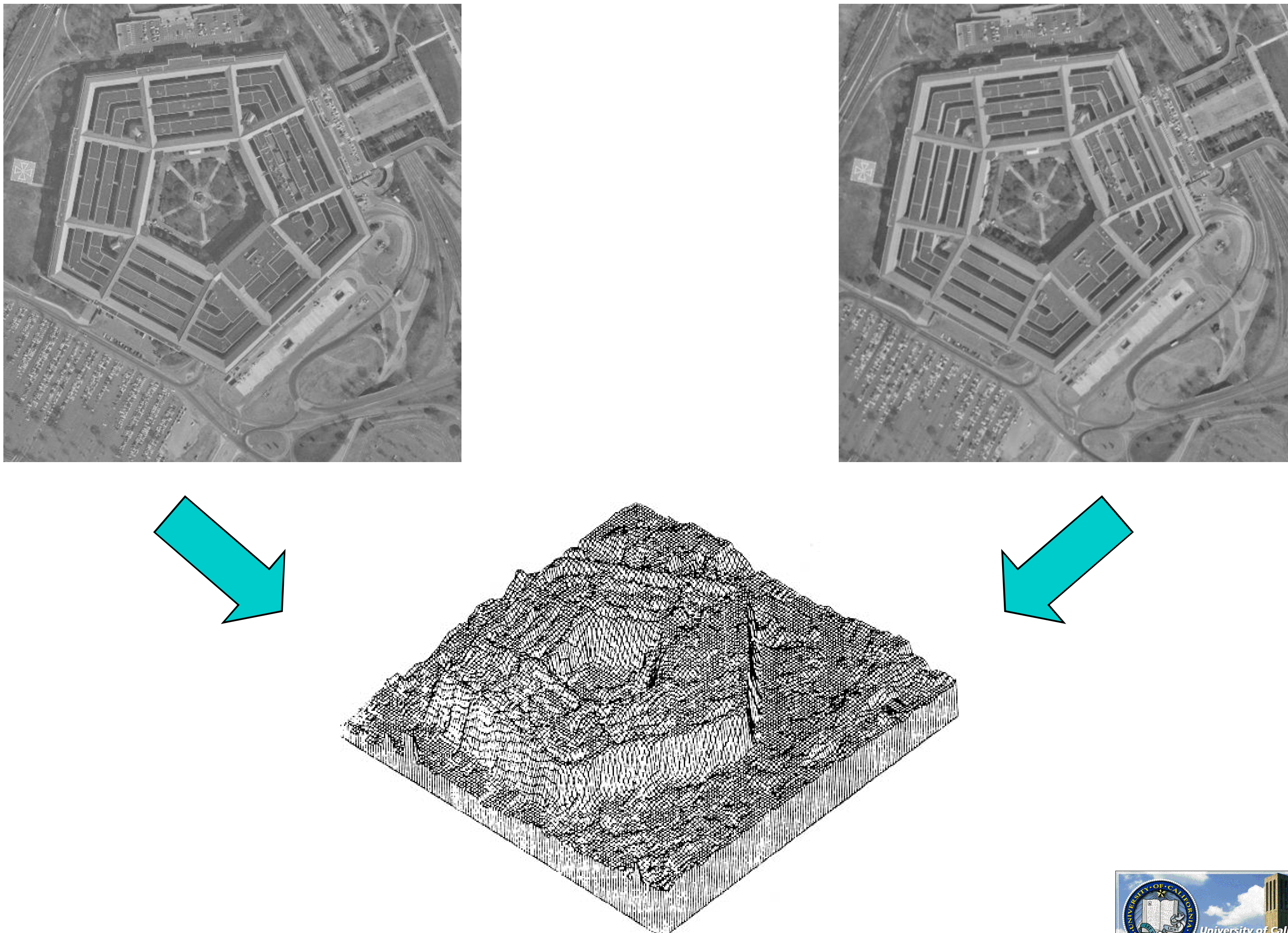


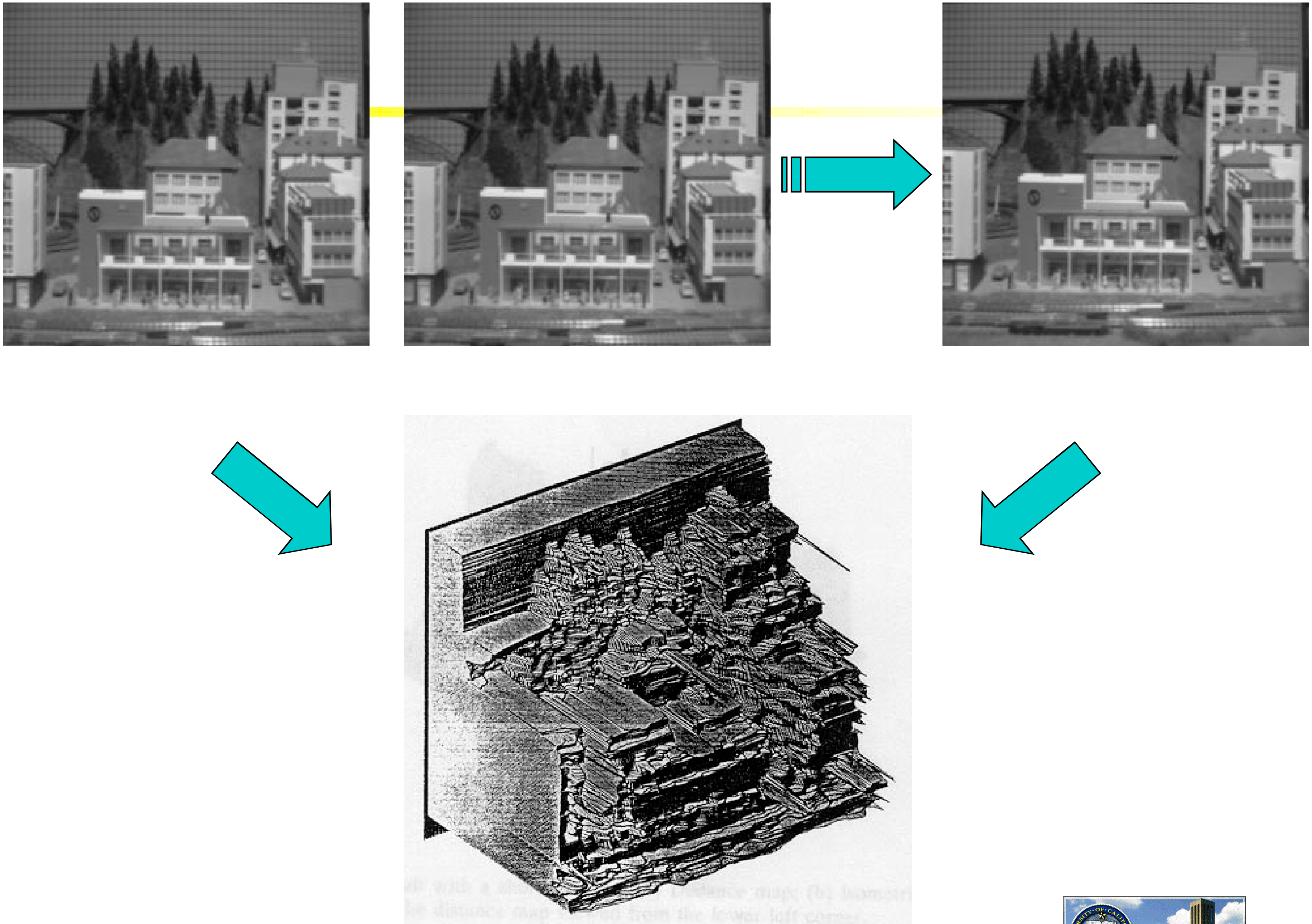
(a)



(b)

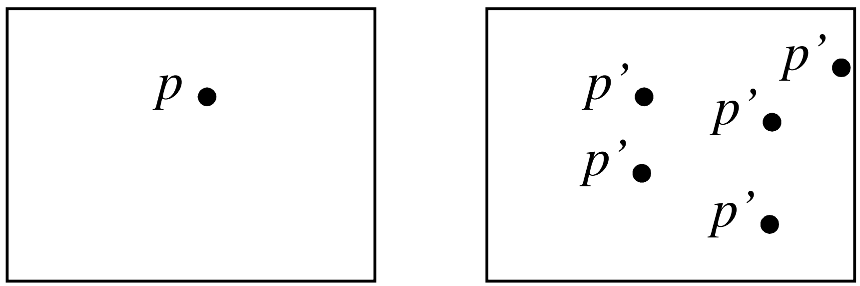
Stereo vision (Stereopsis)





The correspondence problem

- Given a “point” in one image, find the location of that same point in a second image (and maybe third, and fourth, …)



A <u>search</u> problem: Given point ***p*** in the left image, where in the right image should we search for a corresponding point?

Sounds easy, huh?

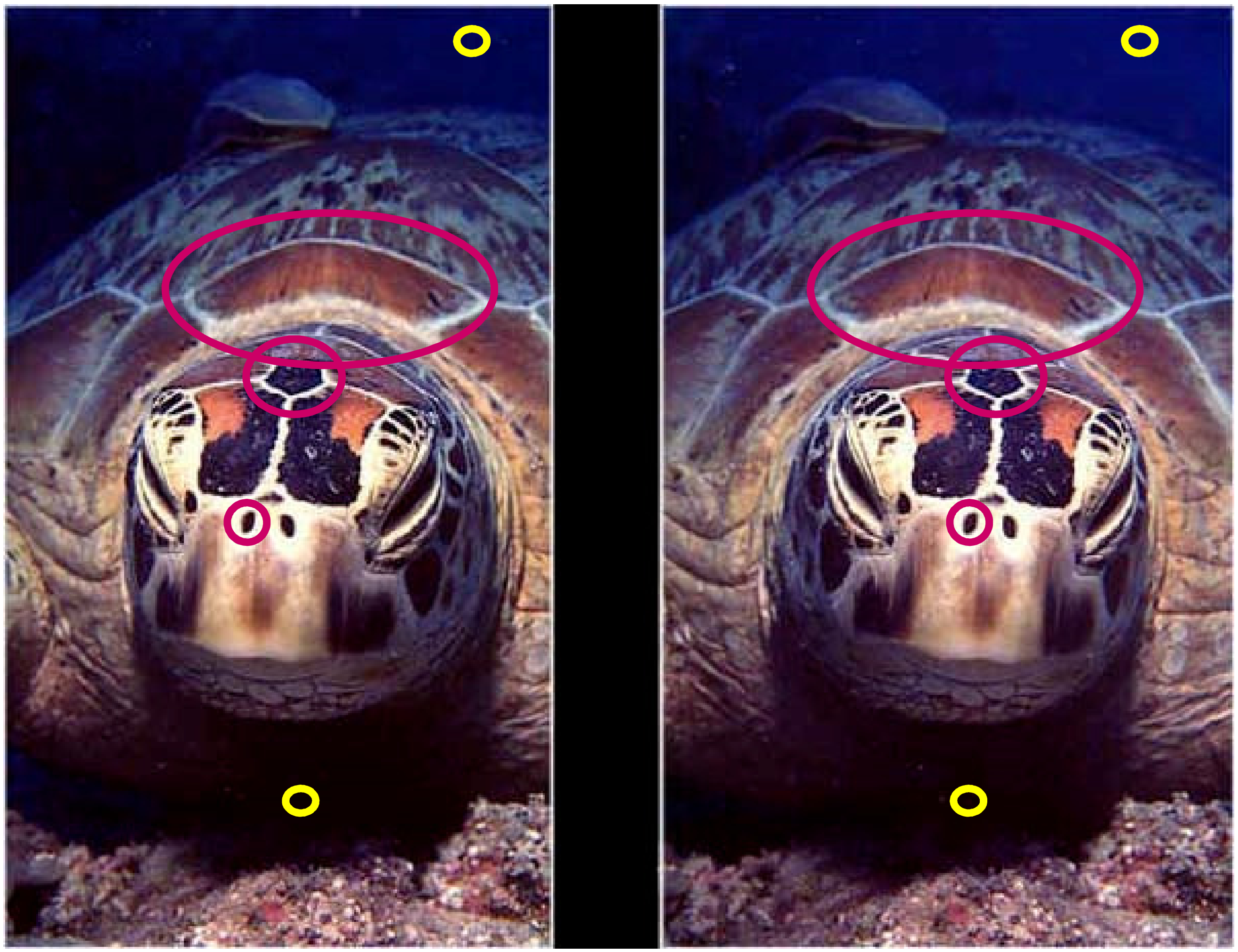
Correspondence example

Left image

Right image

- *What is a point?*
- *How do we compare points in different images? (Similarity measure)*

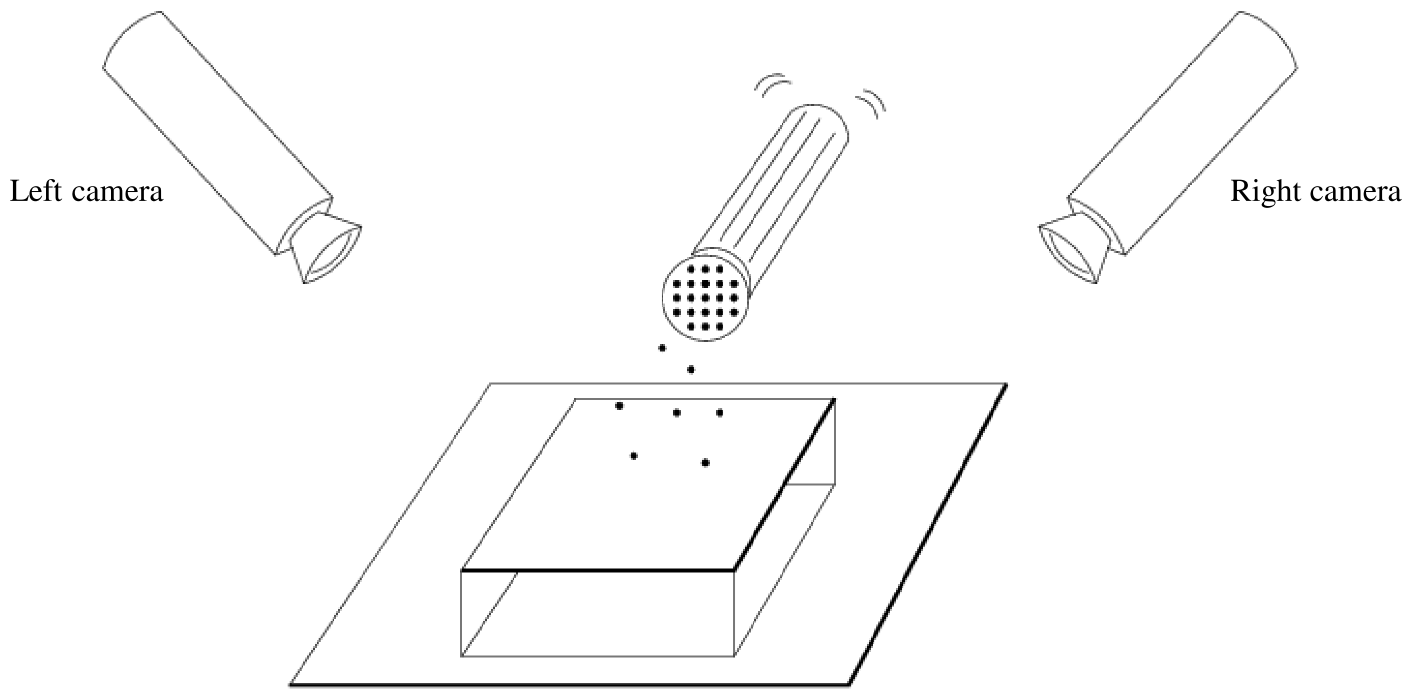
Correspondence example



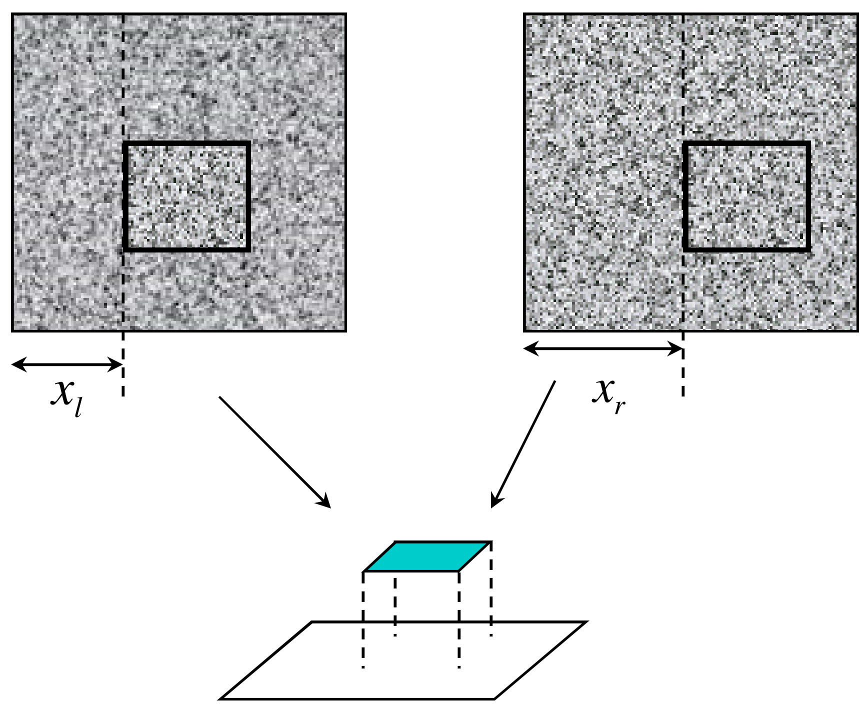
Right image　　　　Left image

Random dot stereograms

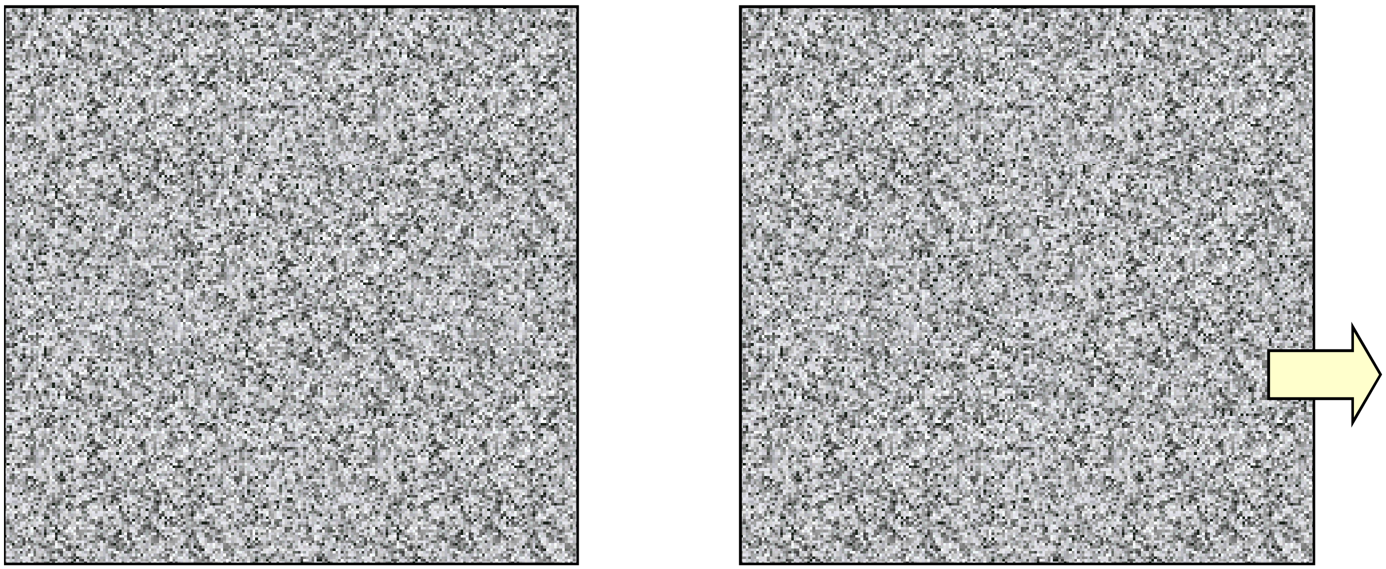
- "High-level" correspondence (recognition) is not always required in order to see depth
- Existence proof: random-dot stereograms

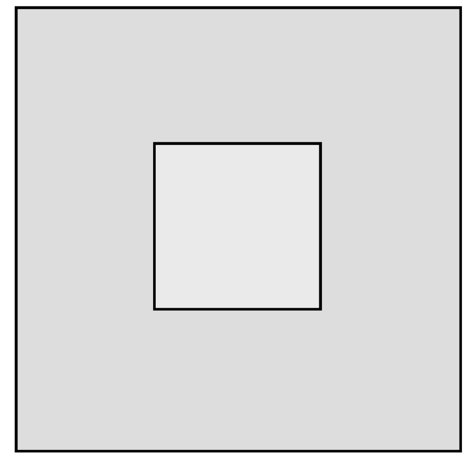


- Q: What features to select?
- A: Intuitively, *unique* and *invariant* (e.g., vehicles, computers)
- Q: Do we need segmentation, recognition, etc. before attempting stereo matching?
- A: *No!*



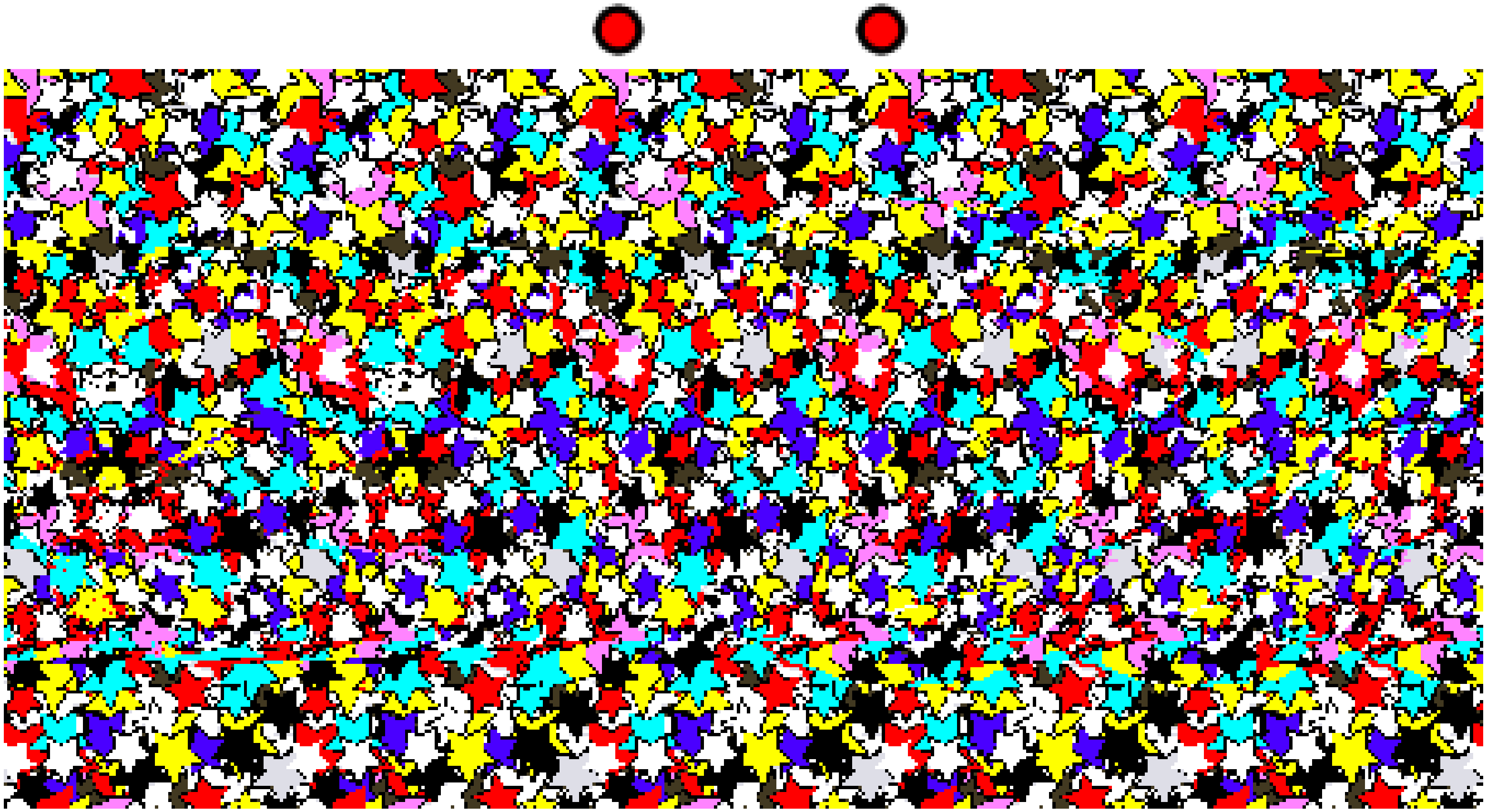
A Random Dot Stereogram



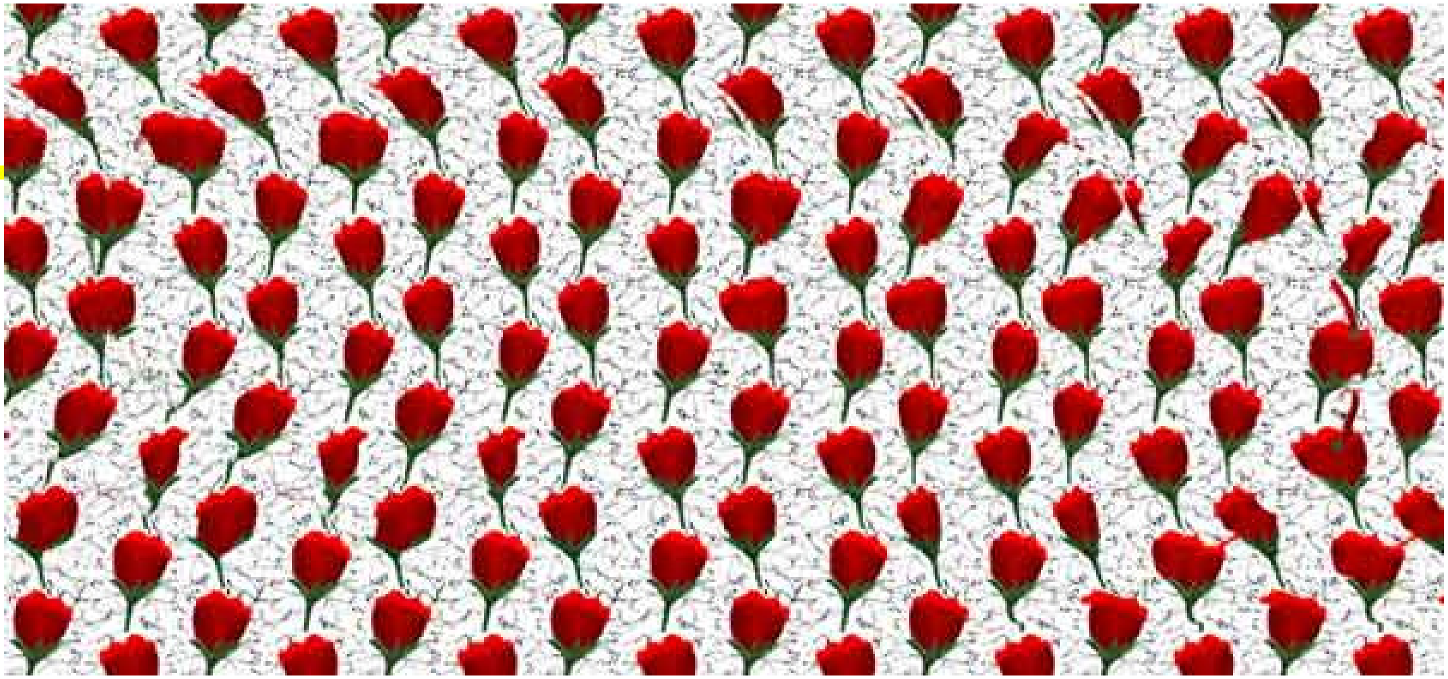


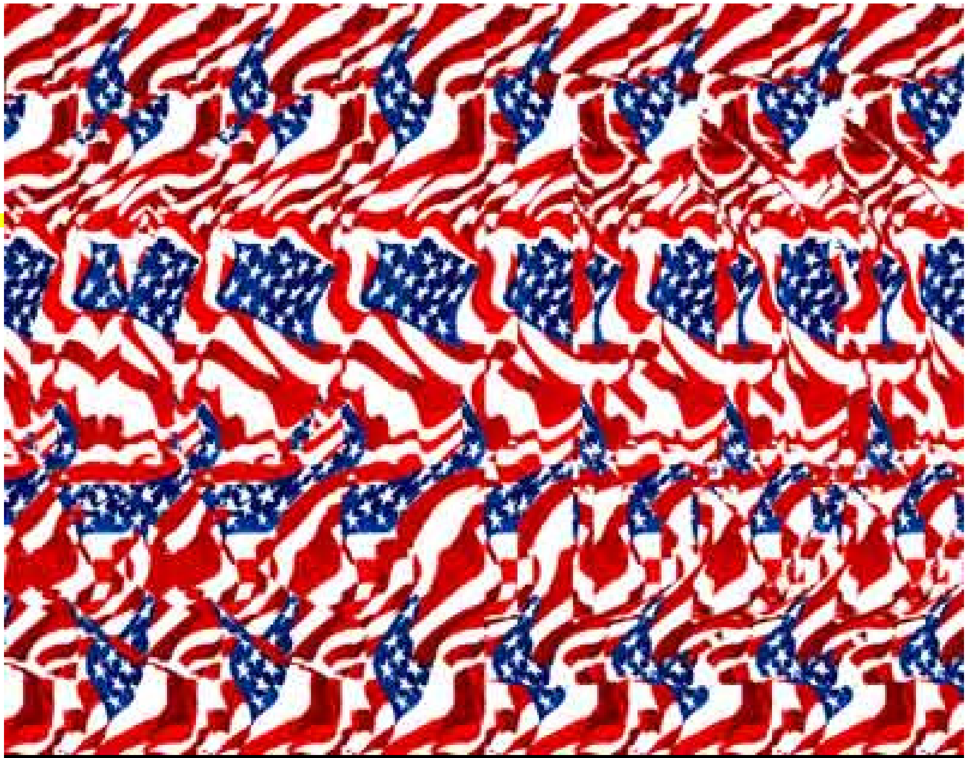
Depth image

Magic Eye Images



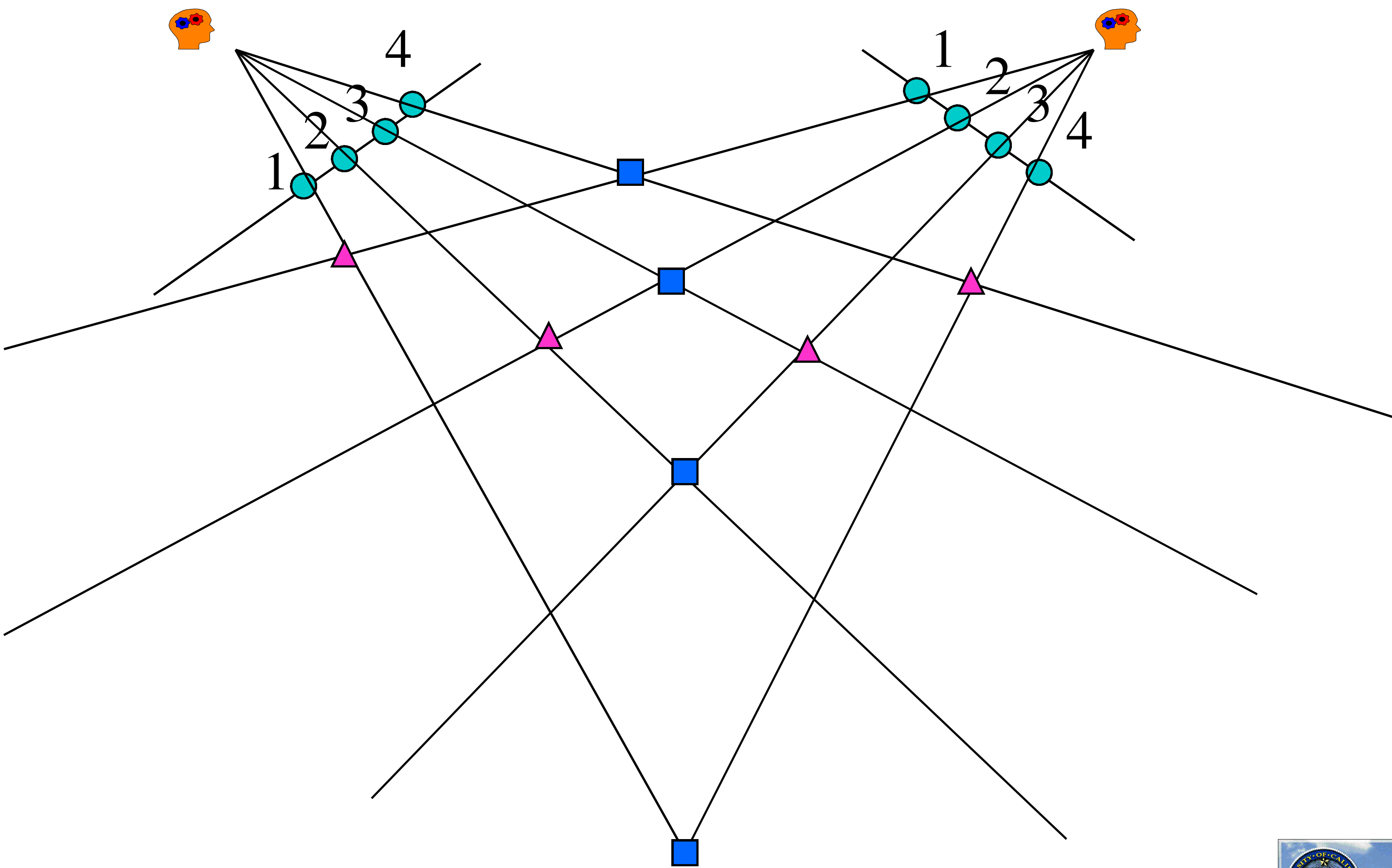
Answer: Saturn





Single image stereograms

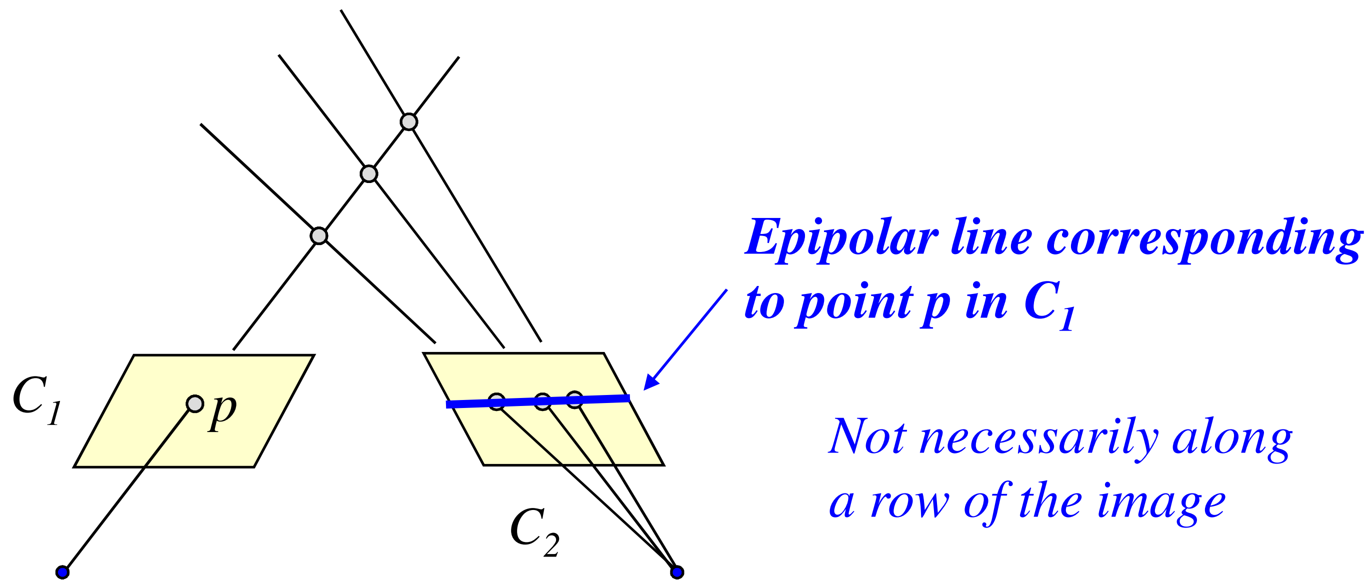
Difficulty in Stereo Correspondence



Difficulty (cont.)

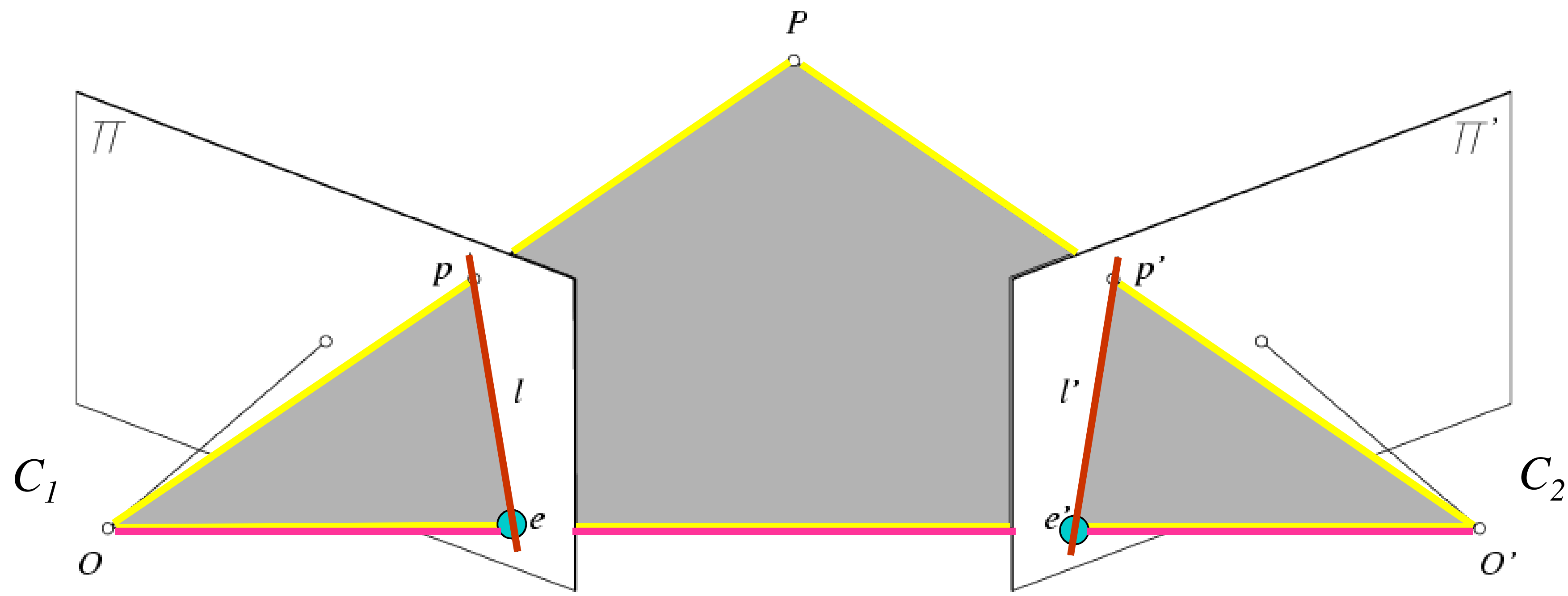
- Multiple matches are always likely
- Simple features (e.g., black dots)
  - large number of potential matches
  - precise disparity
- Complex features (e.g., polygons)
  - small number of potential matches
  - less precise disparity

Two-view geometry



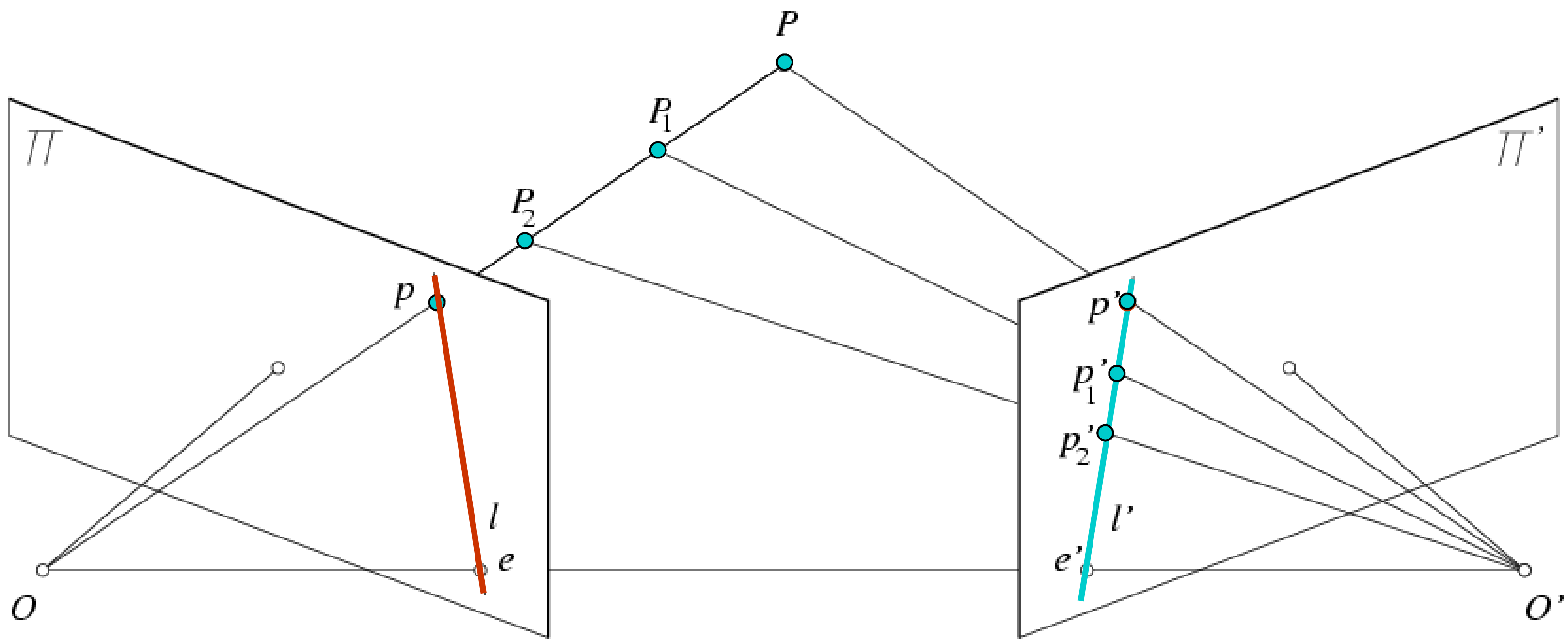
- The ***epipolar geometry*** is defined by the origins of the camera coordinate frames, the scene point *P*, and the locations of the image planes

Epipolar geometry



- Epipolar Plane
- Epipolar Lines
- Epipoles
- Baseline

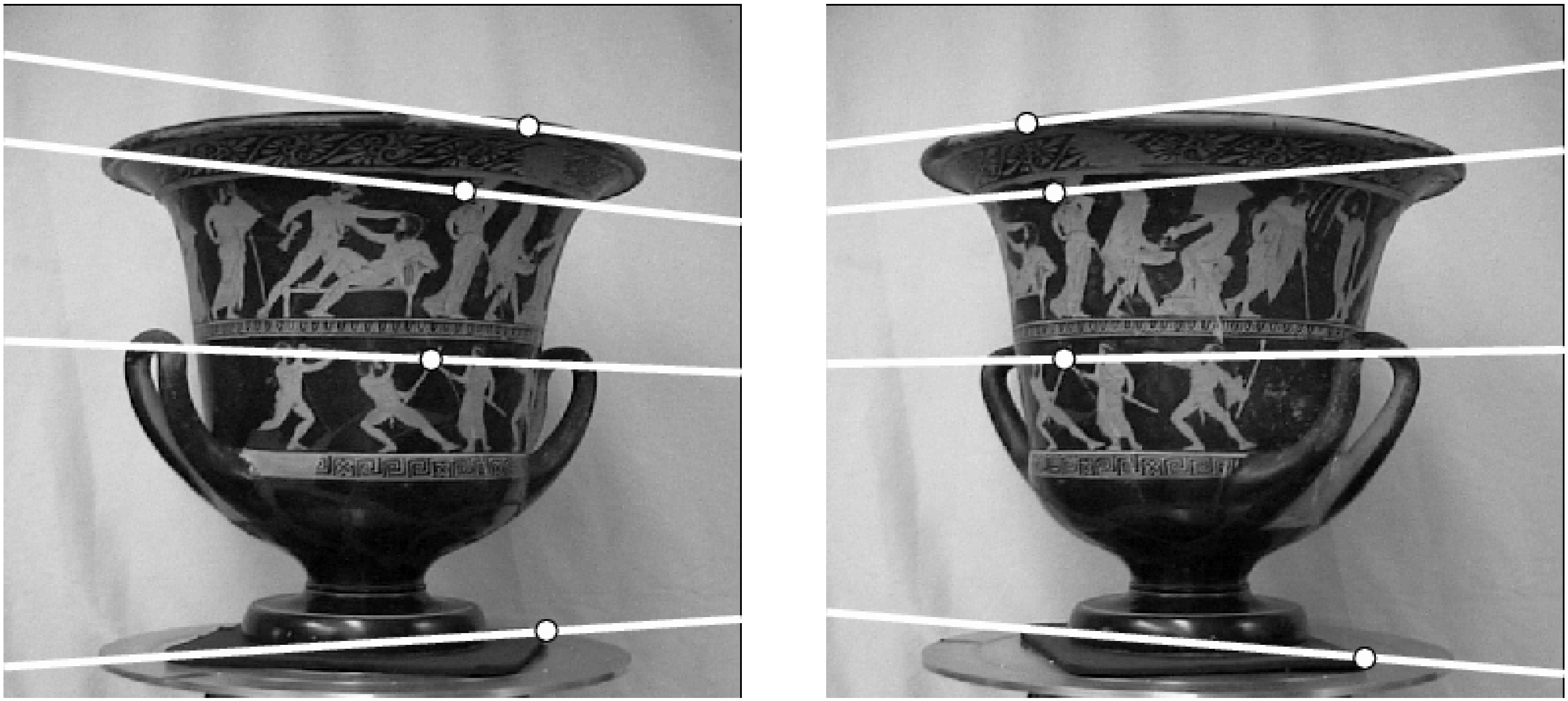
Epipolar constraint



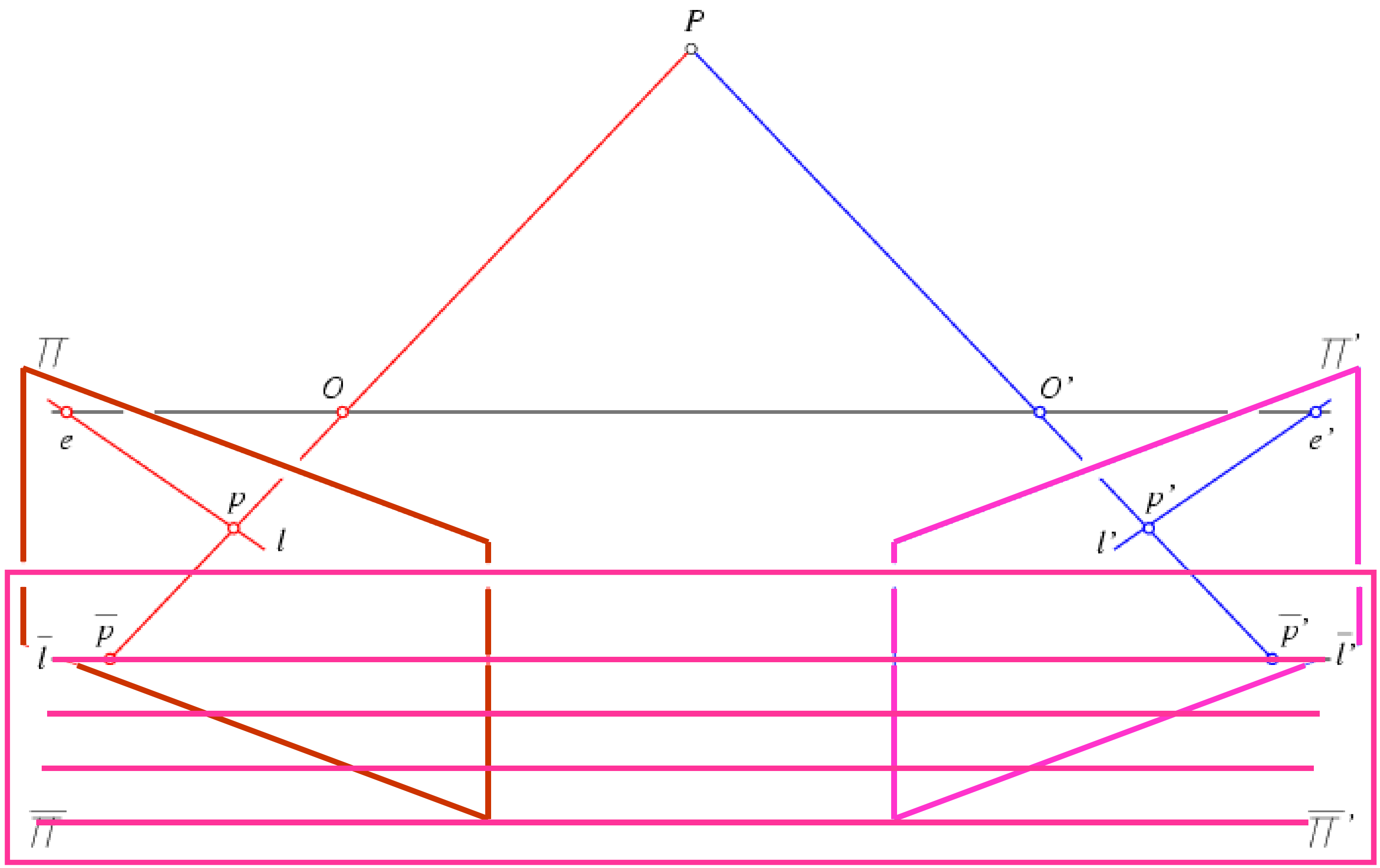
- Potential matches for *p* have to lie on the corresponding epipolar line *l'*

- Potential matches for *p'* have to lie on the corresponding epipolar line *l*

Epipolar lines example

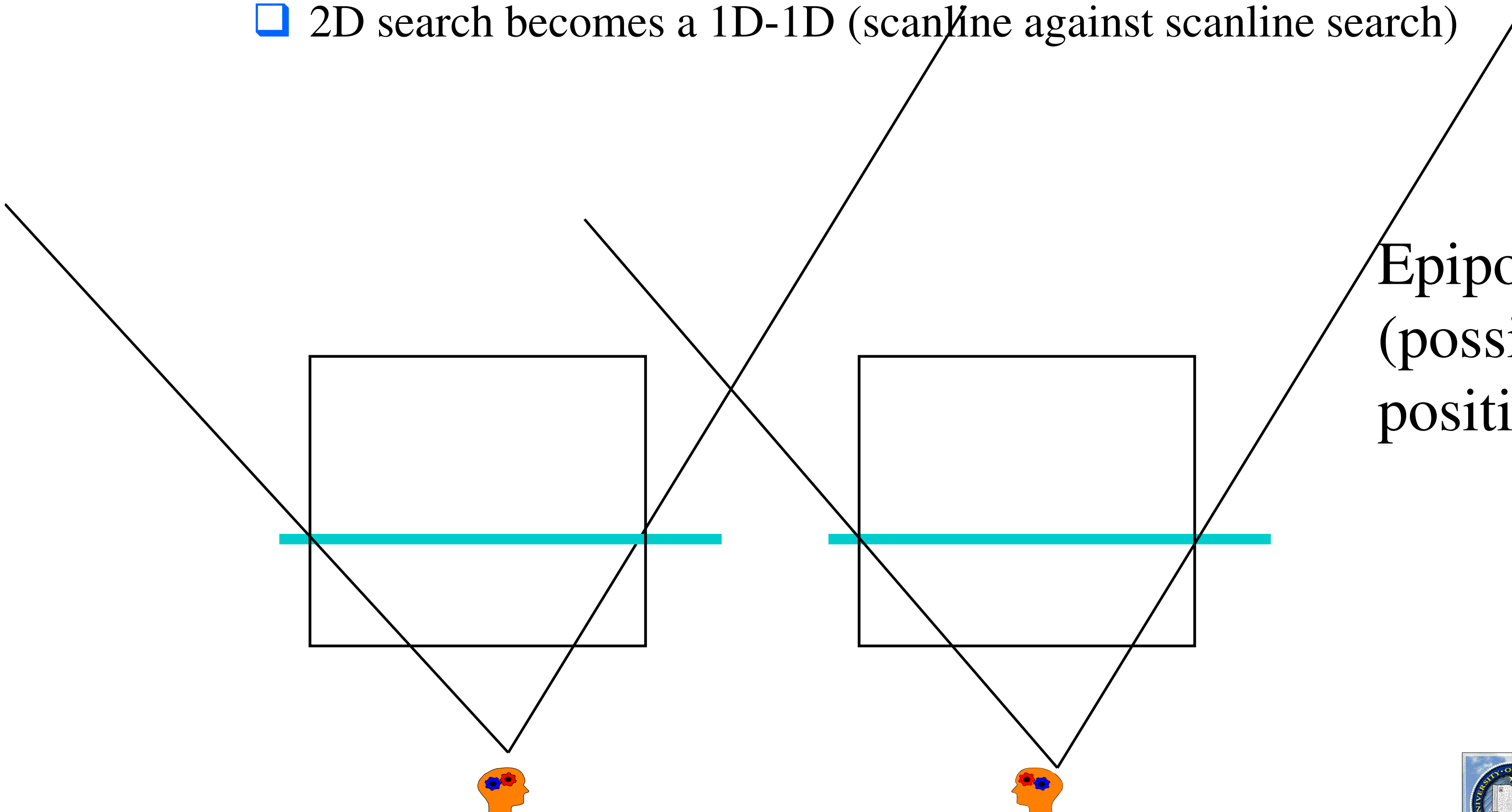


Rectification example



Simple Stereo Correspondence

- *Epipolar constraint for "same image plane" configuration is very simple*
  - Scan line == Epipolar line
  - 2D search becomes a 1D-1D (scanline against scanline search)



Epipolar line (possible match position)

"Standard" Stereo Algorithms

- Assume that images are in simple configuration
- Corresponding scan lines become epipolar lines
- Search can be performed as separate 1D-1D (scanline against scanline) problem
- Many algorithms exist, we describe below three of them

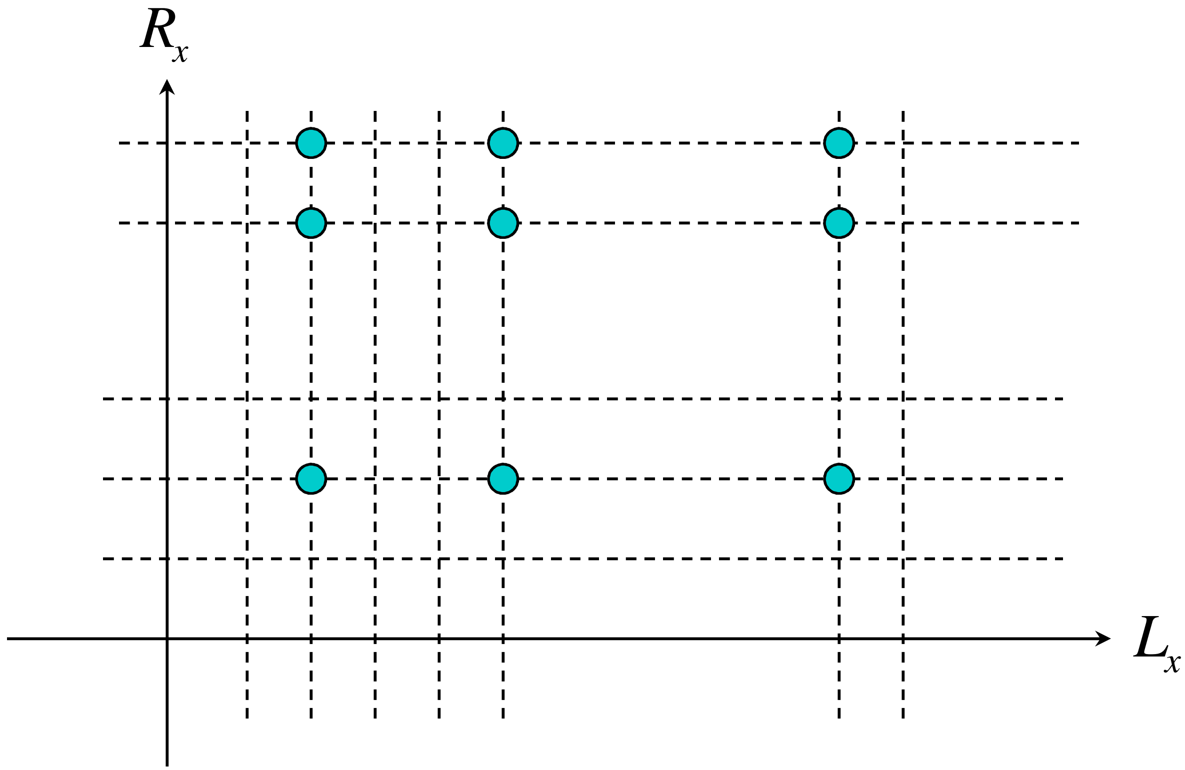
Constraints (cont.)

- *Compatibility*
  - Similar appearance or physical properties (e.g., black dots match black dots)
- *Uniqueness*
  - Projection from 3D to 2D is unique (e.g., one black dot matches *at most* one black dot)
- *Continuity*
  - 3D structures are not random (adjacent dots should have adjacent matches, or similar disparity values)

Marr's algorithm

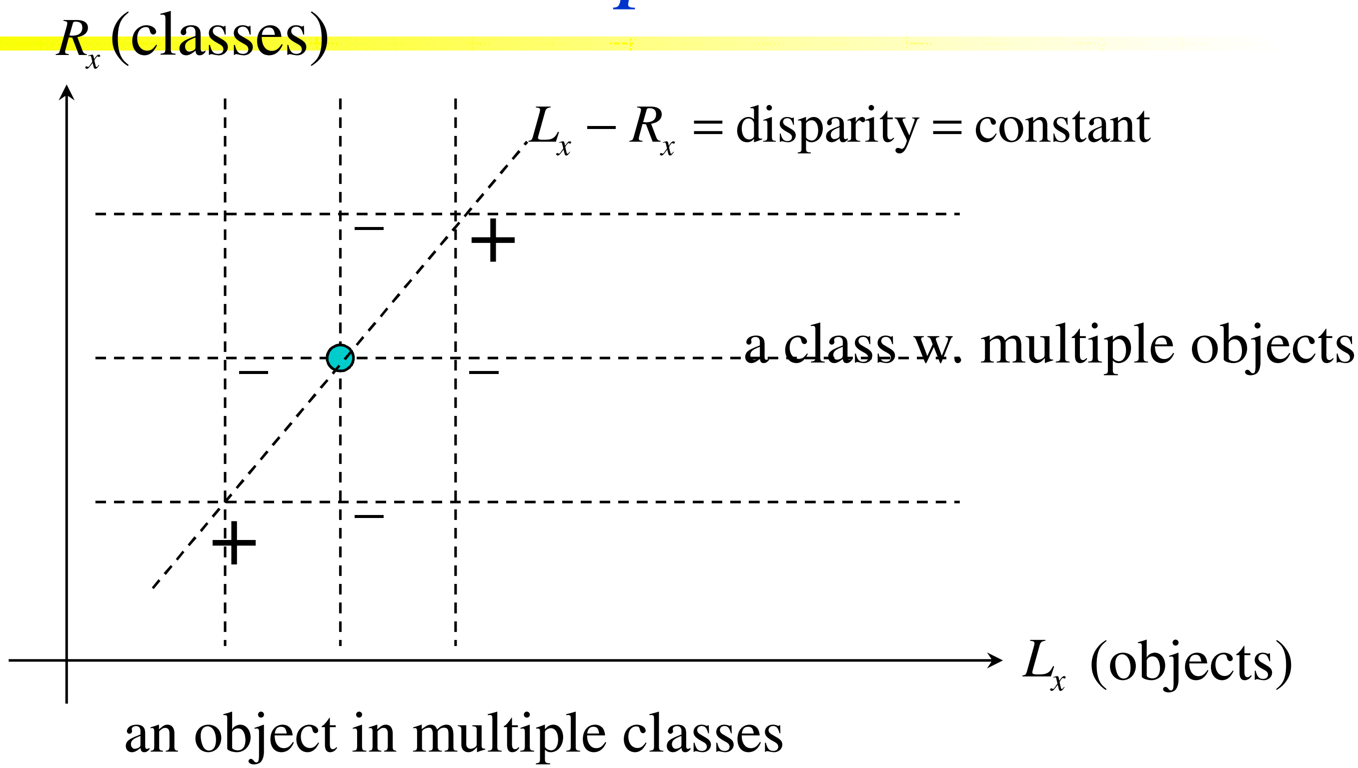
- Based on Relaxation
  - dots in the left images are objects
  - dots in the right images are classes
  - objects should belong in no more than one classes (compatibility and uniqueness)
  - neighboring objects have neighboring classes (continuity)

Initialization



Place a 1 where there is a match of black dots

Update



$$C_{x,y,d}^{(n+1)} = \sigma\left\{ \sum_{x',y',d' \in S(x,y,d)} C_{x',y',d'}^{(n)} - \xi \sum_{x',y',d' \in O(x,y,d)} C_{x',y',d'}^{(n)} + C_{x,y,d}^{(0)} \right\}$$

n: iteration number

$S(x, y, d)$: local excitatory neighborhood

$O(x, y, d)$: local excitatory neighborhood

ξ: inhibition constant

σ: threshold function

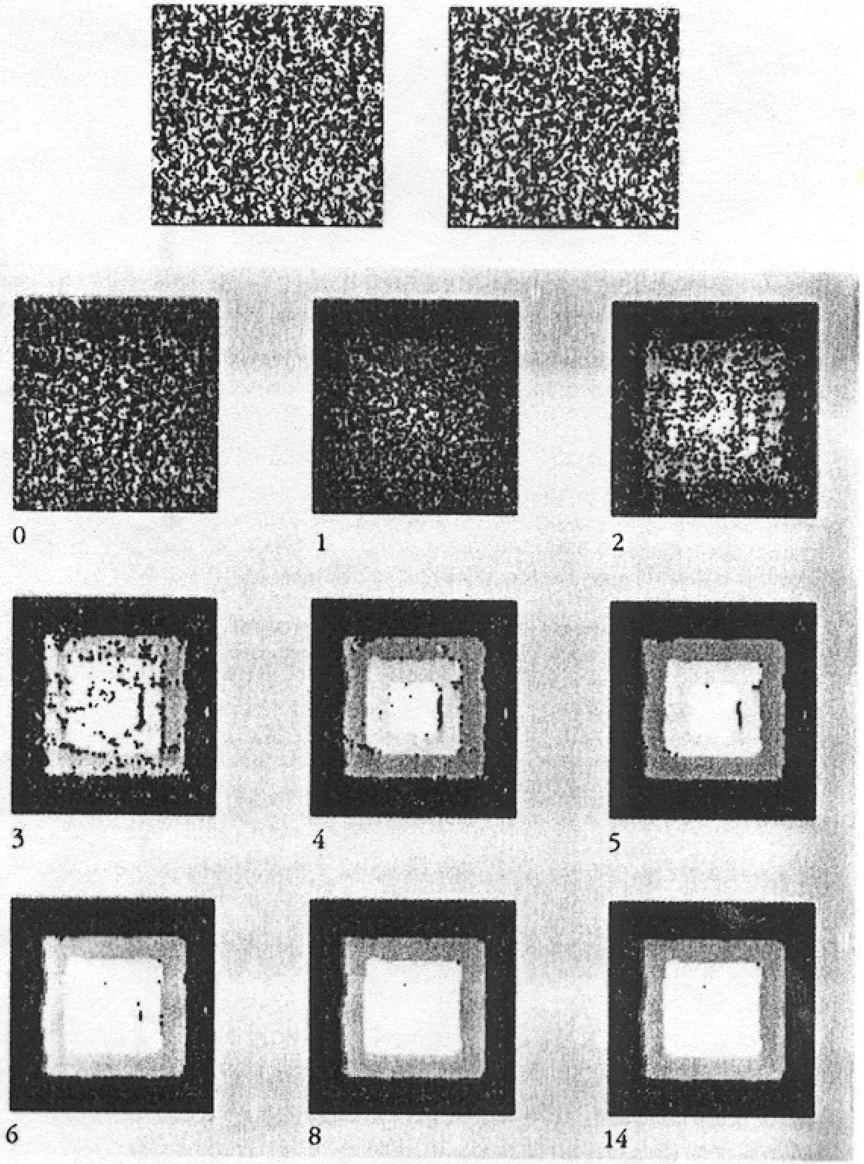
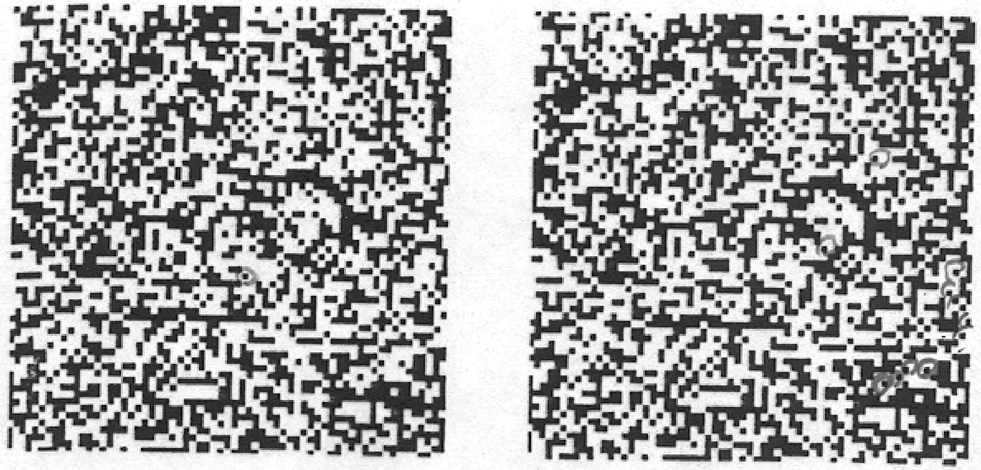
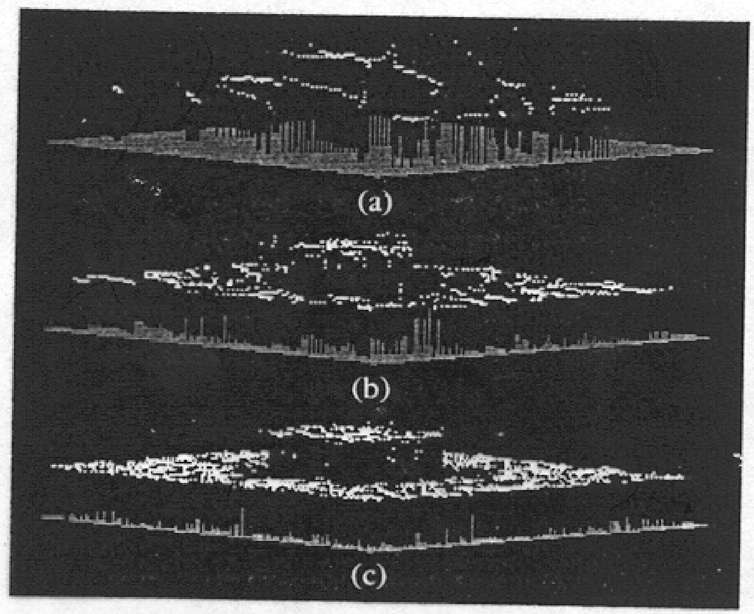


Figure 3–7. The decoding of a random-dot stereogram pair by the cooperative algorithm described in the text. The stereogram appears at the top, and the initial state of the network, which includes all possible matches within the prescribed disparity range, is labeled 0. The algorithm runs through a number of iterations, as shown, and gradually the structure is revealed. The different shades of gray represent different disparity values.



Left Right



Optimization Algorithm

- Goal: to find a function that satisfies the *compatibility, uniqueness and continuity* constraints
- Q: What function?
- A:
  - in 3D $Z(x,y)$ (depth)
  - in 2D $d(x,y)$ (disparity)
  - in either case, *uniqueness* constraint is implicitly satisfied

Compatibility

- Q: How about compatibility?
- A: Similar intensity (brightness, pattern, etc.) at matched points

$$I^l{}_{(left\ feature)} = I^r{}_{(right\ feature)}$$

$$I^l(x + \frac{1}{2}d, y) = I^r(x - \frac{1}{2}d, y)$$

Continuity

- Q: How about *continuity*?
- A: Local variation in disparity should be small

$$\left(\frac{\partial d}{\partial x}\right)^2 \cong 0$$

Mathematically

$$\text{minimize } E = \int_x (I^l(x+\frac{1}{2}d) - I^r(x-\frac{1}{2}d))^2 + \lambda[(\frac{\partial d}{\partial x})^2 + (\frac{\partial d}{\partial y})^2]$$

Discrete case

$$E = \sum_i (I^l(x_i+\frac{1}{2}d_i) - I^r(x_i-\frac{1}{2}d_i))^2 + \lambda(d_{i+1}-d_i)^2$$

$$\frac{\partial E}{\partial d_k} = (I^l(x_k+\frac{1}{2}d_k) - I^r(x_k-\frac{1}{2}d_k))(\frac{\partial I^l}{\partial x} + \frac{\partial I^r}{\partial x}) - 2\lambda[(d_{k+1}-d_k)+(d_{k-1}-d_k)] = 0$$

$$d_k = \overline{d}_k - \frac{1}{\lambda}(I^l(x_k+\frac{1}{2}d_k) - I^r(x_k-\frac{1}{2}d_k))(\frac{\partial I^l}{\partial x} + \frac{\partial I^r}{\partial x})$$

Results

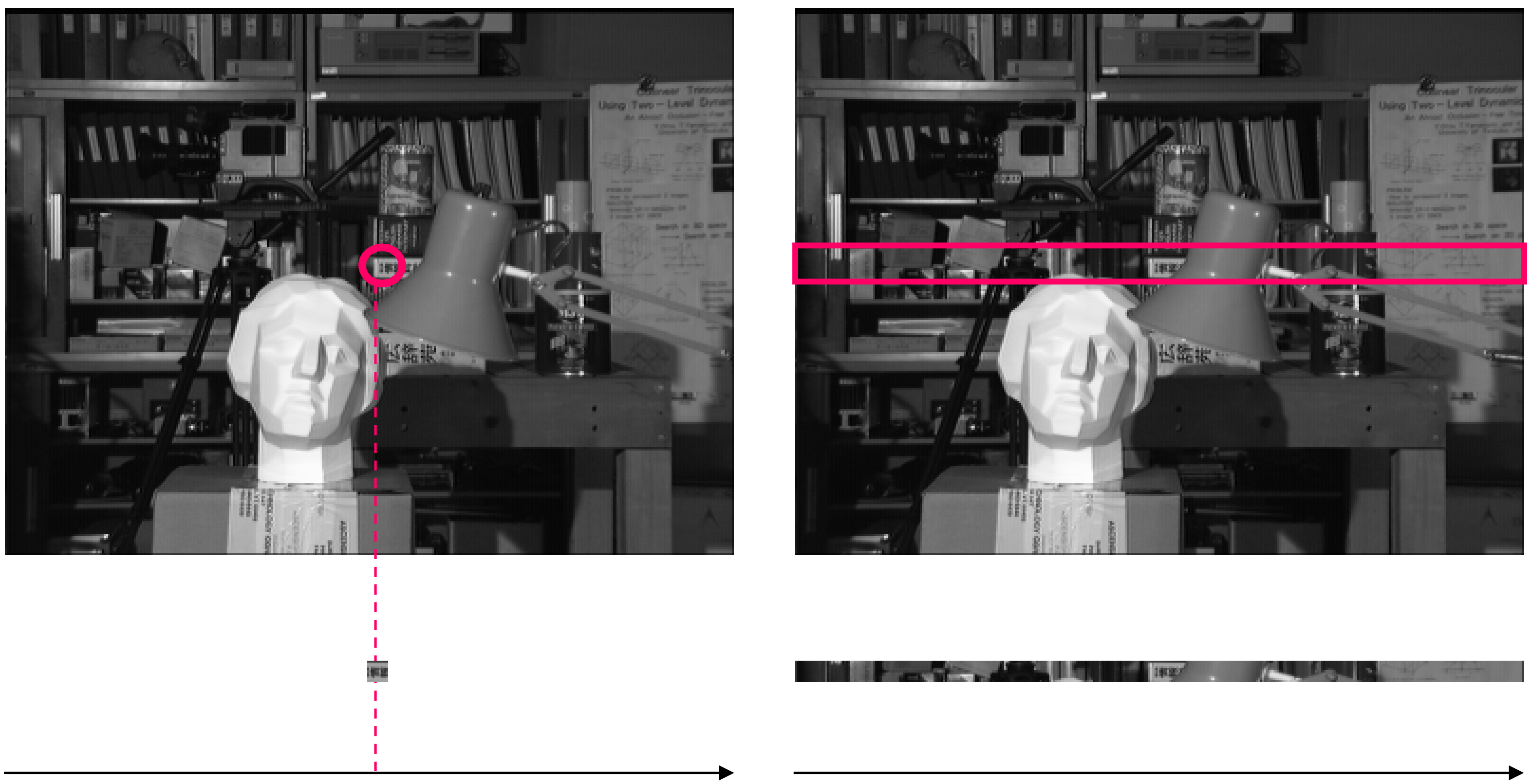
$$d_k = \boxed{\bar{d}_k} - \boxed{\frac{1}{\lambda}}\boxed{\left(I^l\left(x_k + \frac{1}{2}d_k\right) - I^r\left(x_k - \frac{1}{2}d_k\right)\right)}\boxed{\left(\frac{\partial I^l}{\partial x} + \frac{\partial I^r}{\partial x}\right)}$$

- estimate based on smoothness
- how much does the smooth estimate violate similarity constraint
- how much does that matters
- direction for correction (there better be changes in intensity, otherwise, correction will not help reducing matching error)

Stereo matching

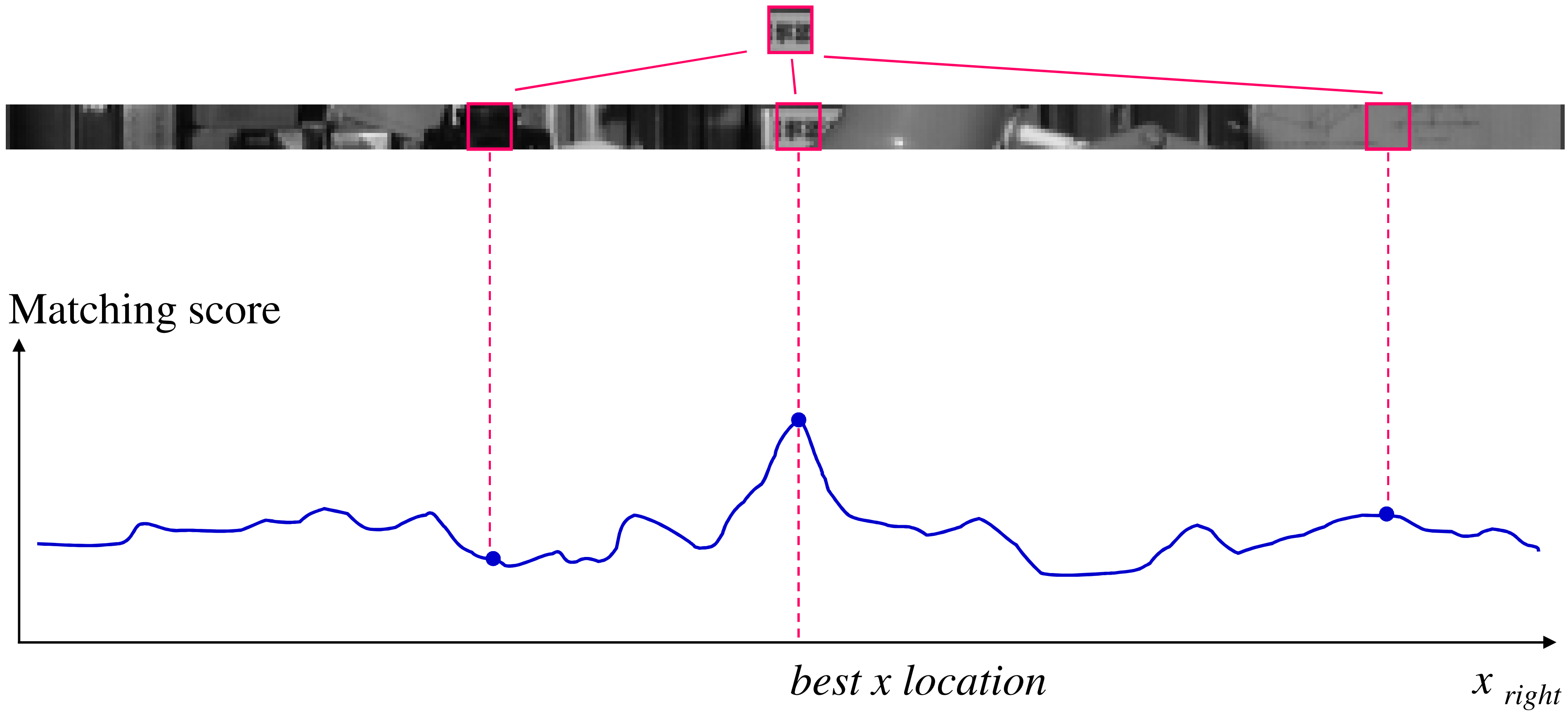
Rectified images

Left Right



x_{left} x_{right}

Matching along epipolar line

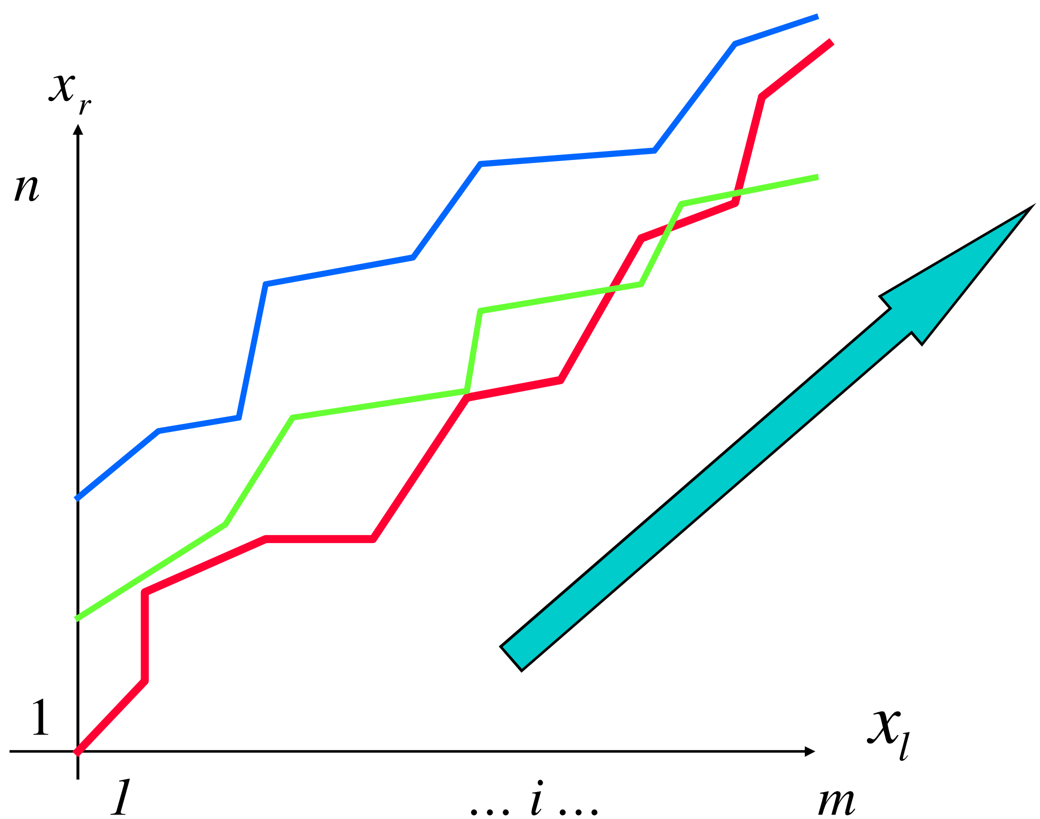


The best match estimates the "disparity" ***δu***

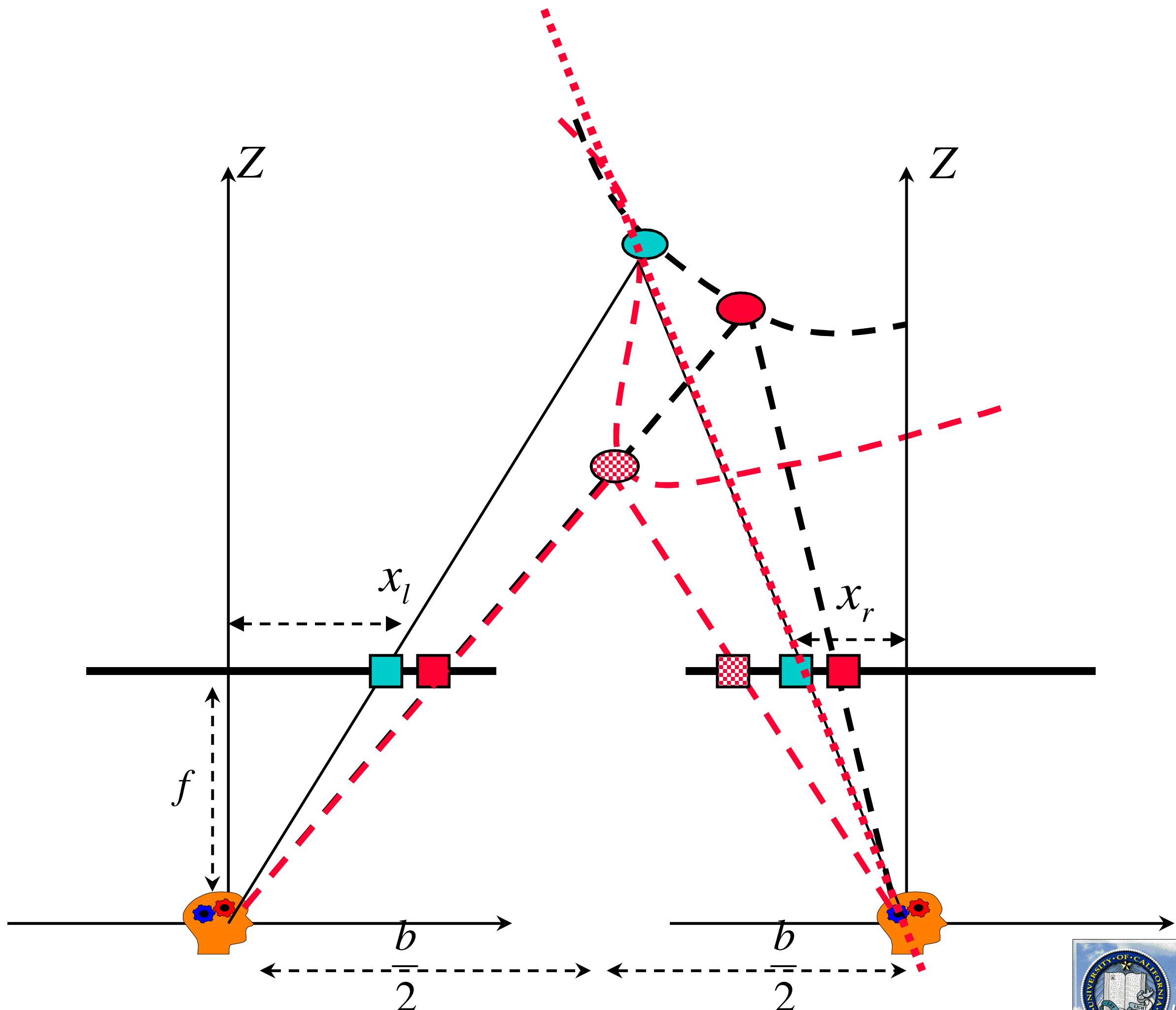
- In this case, horizontal disparity only (since images were rectified)

Dynamic Programming

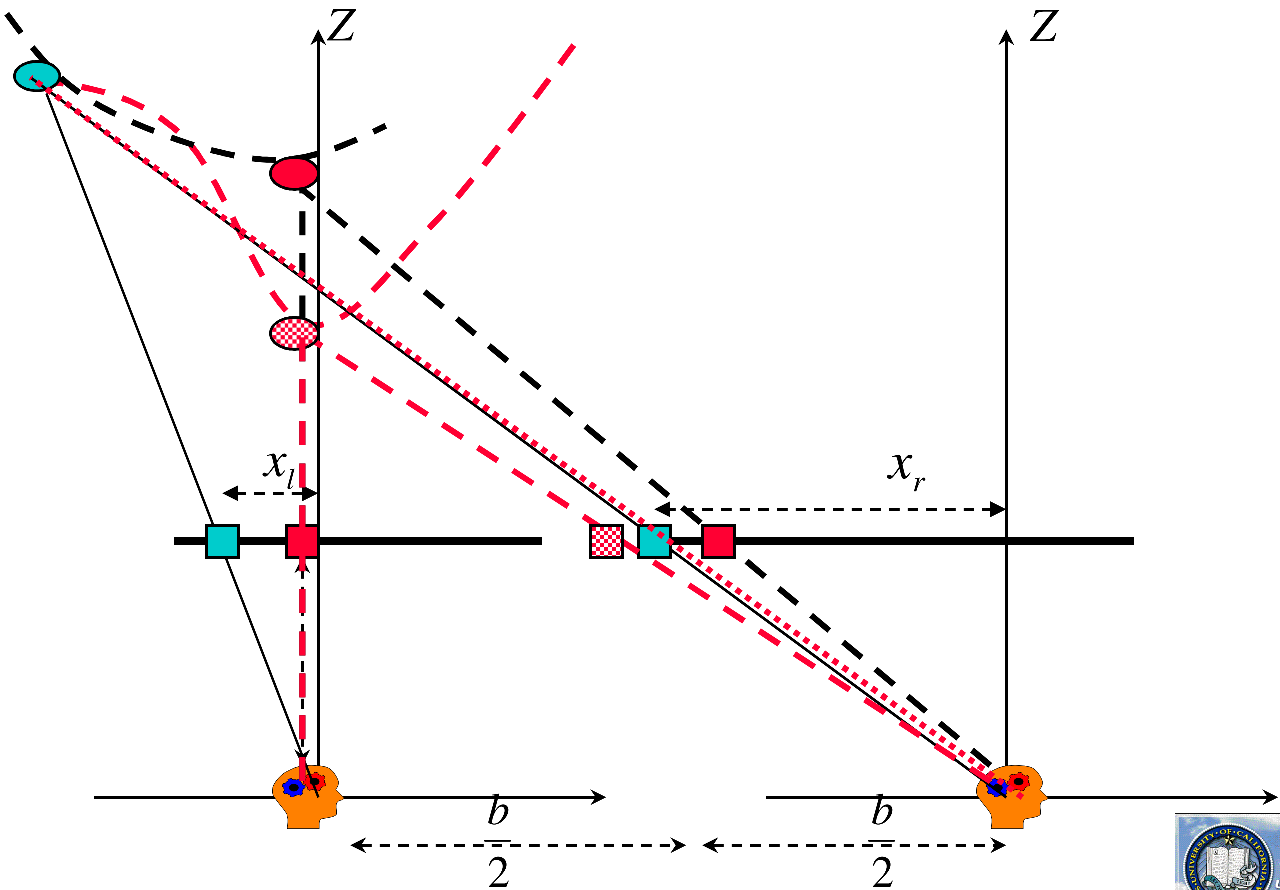
- Finding a path in a 2D matrix representing two corresponding epipolar lines
- Additional constraint – path can go only one way



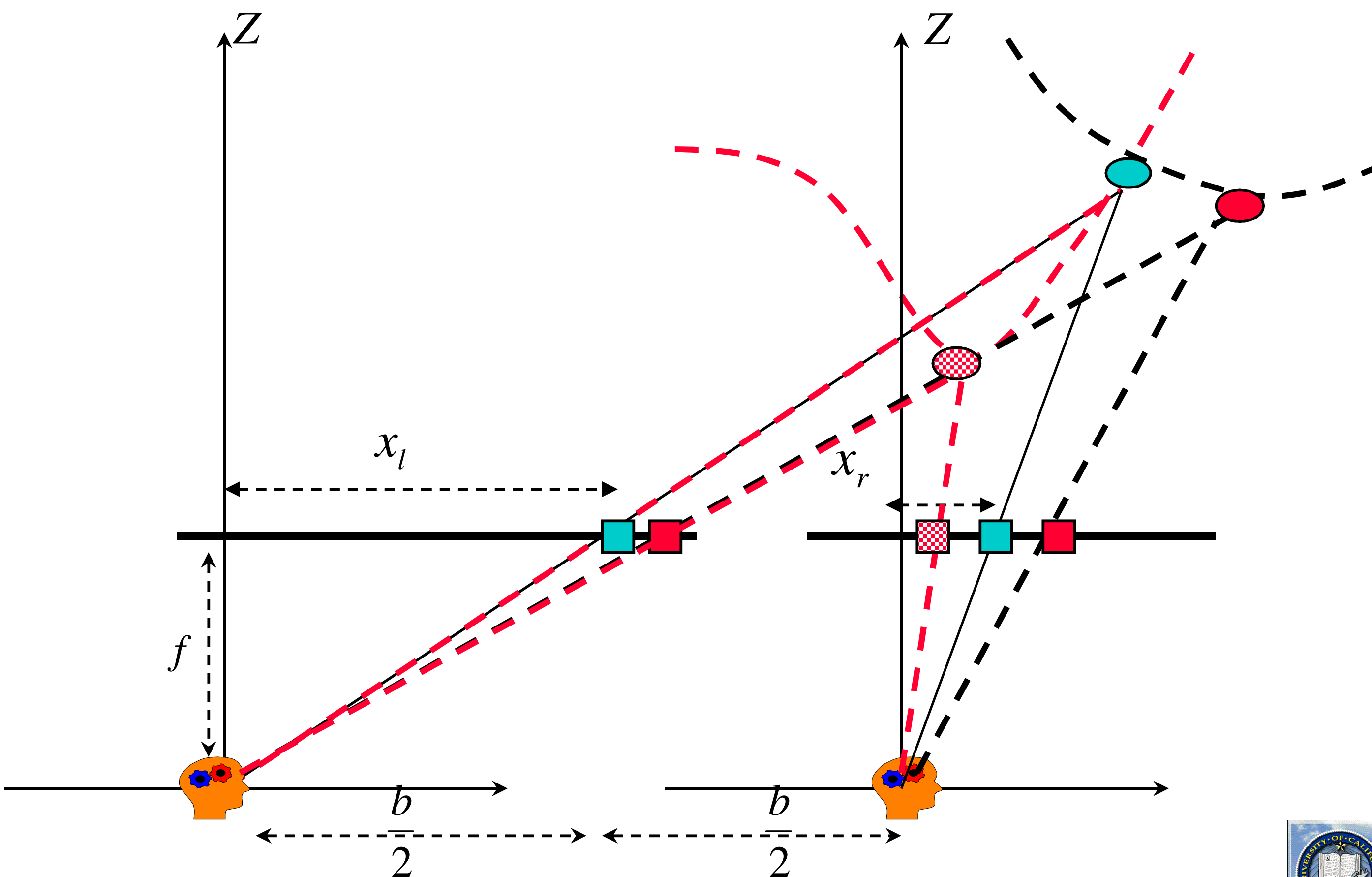
DP Path Constraints



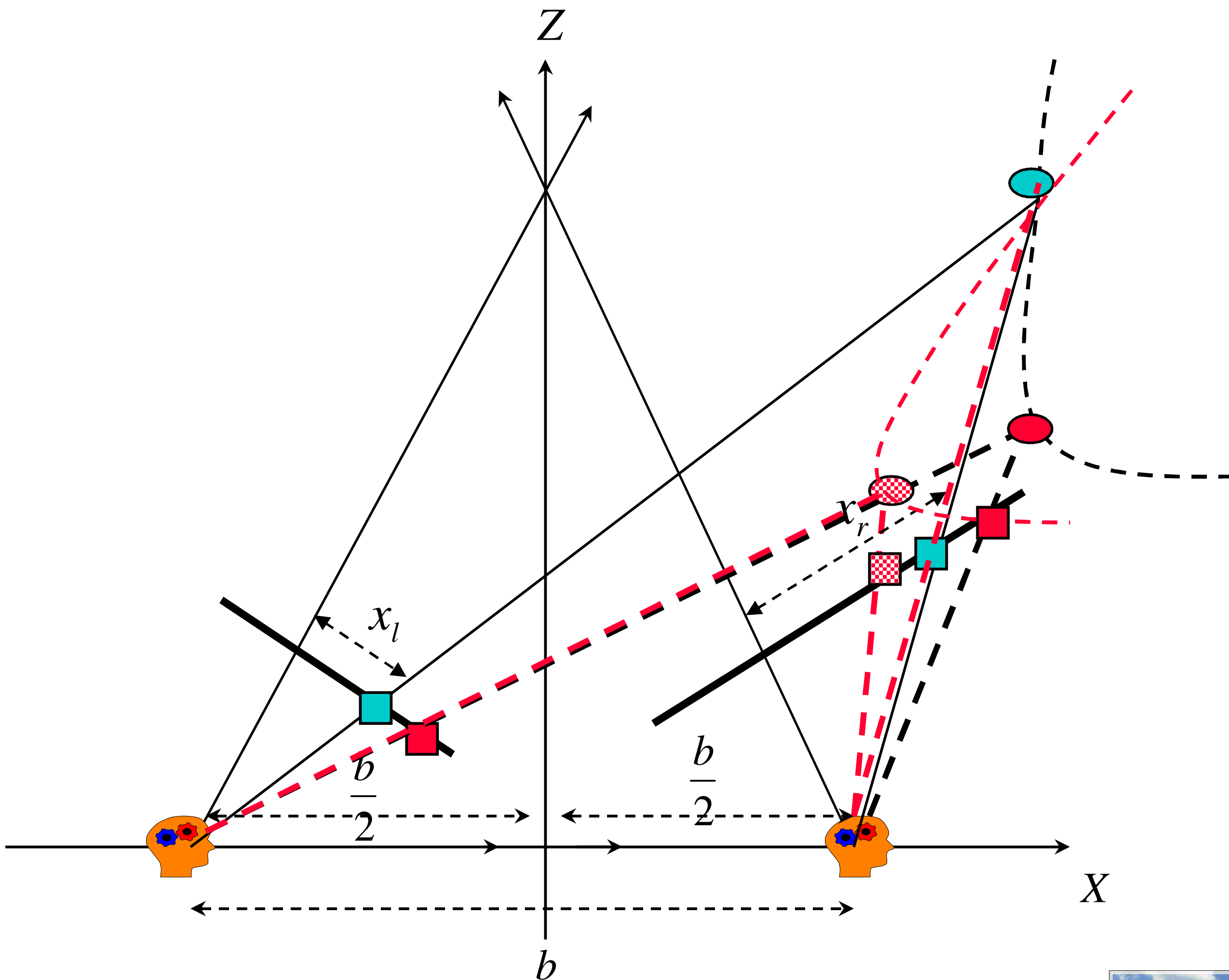
DP Path Constraints (cont)



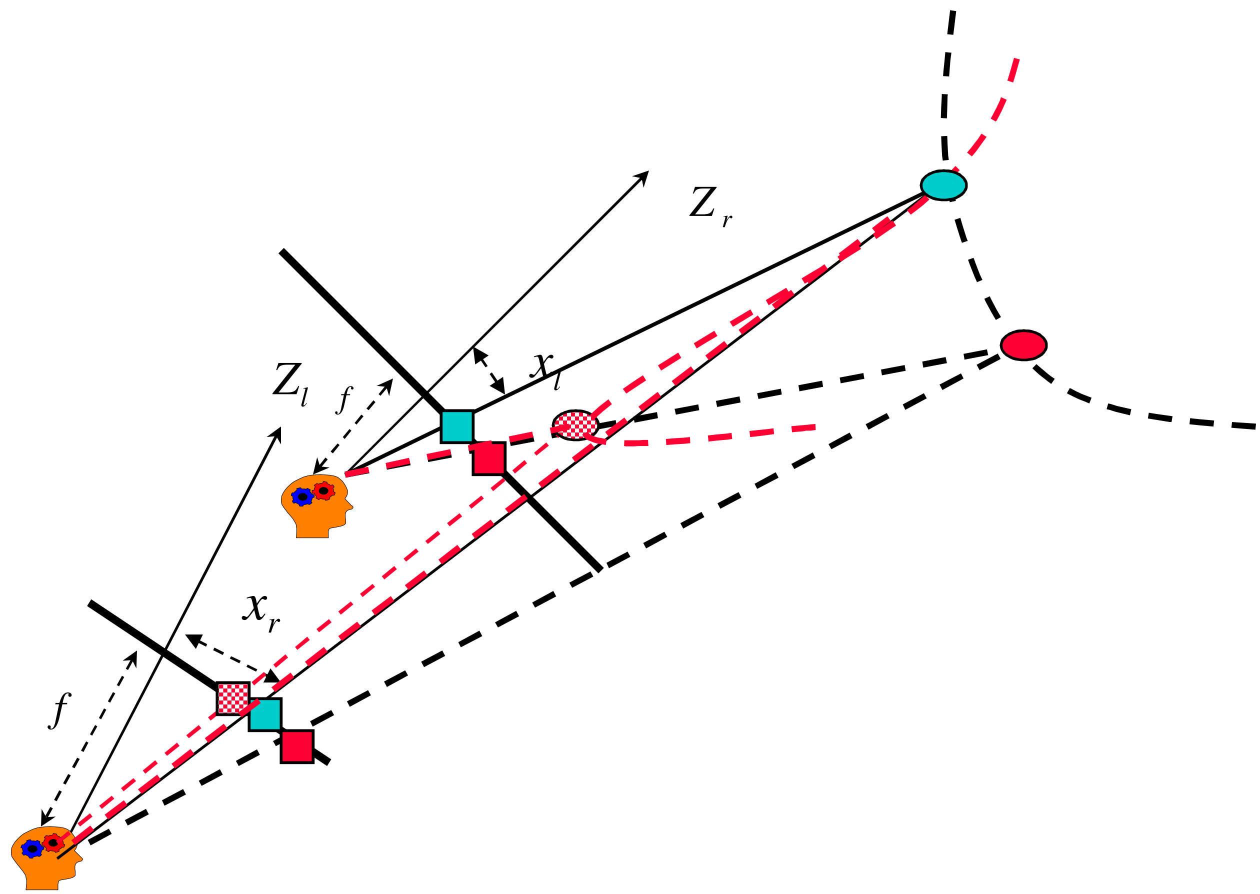
DP Path Constraints (cont)



Valid for Other Stereo Configurations



Valid for Other Stereo Configurations

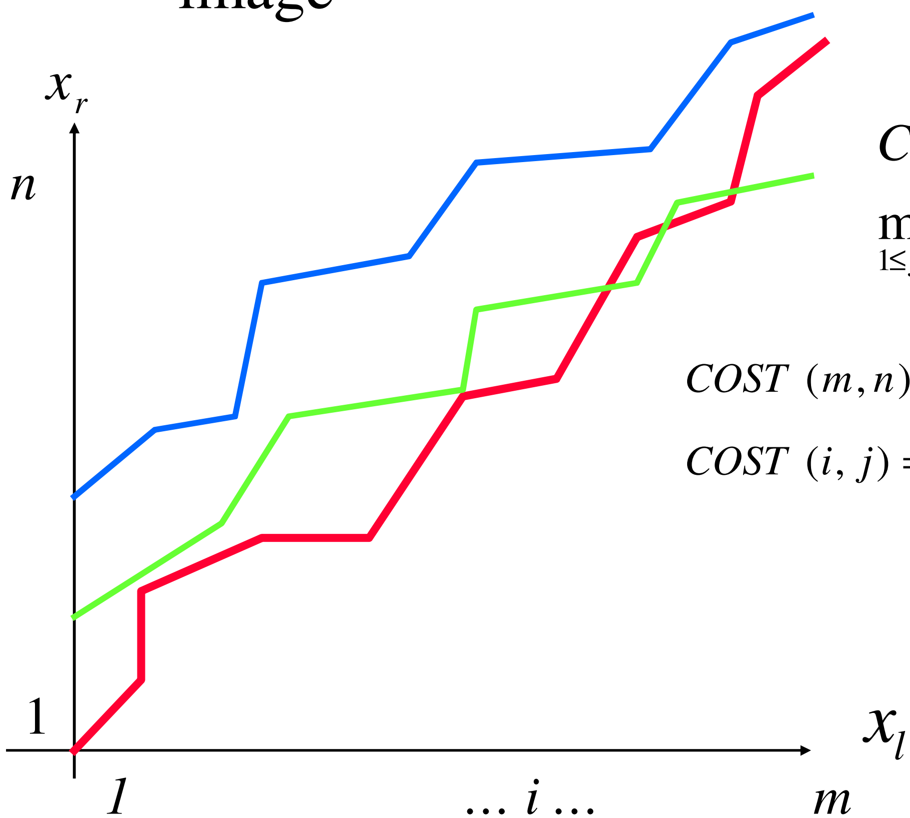


DP Constraints (cont.)

- *Compatibility*
  - Similar appearance or physical properties (e.g., black dots match black dots)
- *Uniqueness (DP Path constraint)*
  - Projection from 3D to 2D is unique (e.g., one black dot matches *at most* one black dot)
  - Path should not go vertical or horizontal
- *Continuity (DP Path constraint)*
  - Path should go only one way (from lower left to upper right)
    - If x_l matches x_r
    - $x_l + 1$ matches $x_r + 1 + d$ ($d >= 0$) (on a discrete grid)
  - Change of d should be smooth

Recursion

- *COST(m,n):* total cost of matching *m* points on left image with *n* points on the right image
- *C(i,j):* matching pixel *i* in left image with pixel *j* in right image



$$COST(m,n) = \min[C(m,\phi) + COST(m-1,n), \min_{1\le j\le n} C(m,j) + COST(m-1,j-1)]$$

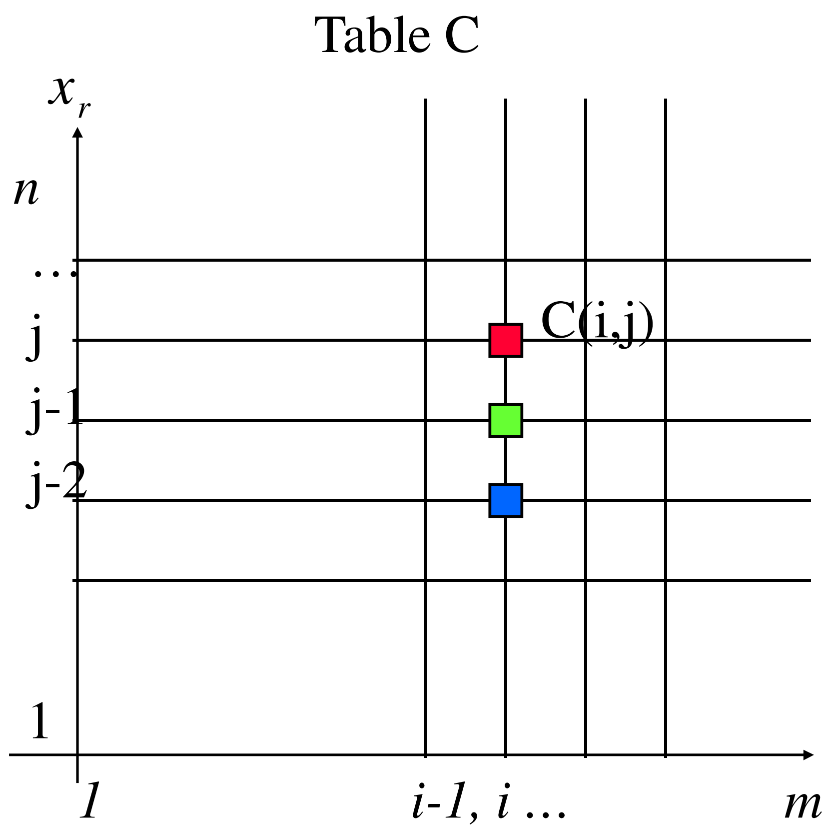
$$COST\ (m,n) = \min_{1\le j\le n+1}[C(m,i) + COST\ (m-1,i-1)],\ \phi = n+1$$

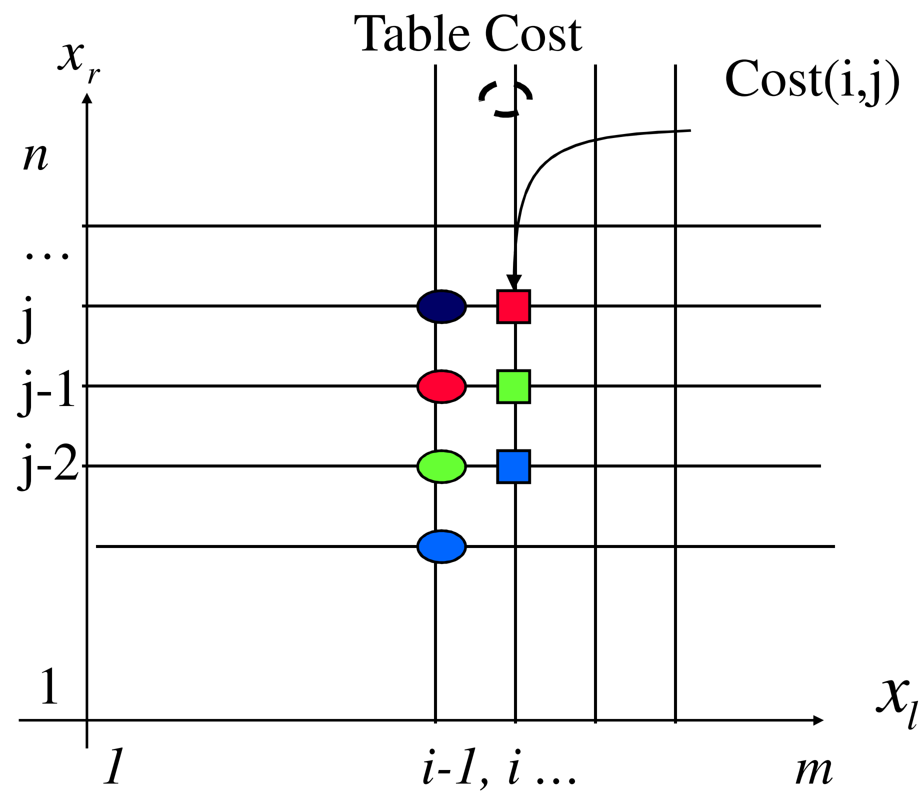
$$COST\ (i,j) = \min_{1\le k\le j+1}[C(i,k) + COST\ (i-1,k-1)],\ \phi = j+1$$

Table Building - Iteration

$$COST(m,n) = \min_{1\le j\le n+1}[C(m,i)+COST(m-1,i-1)], \phi = n+1$$

$$COST(i,j) = \min_{1\le k\le j+1}[C(i,k)+COST(i-1,k-1)], \phi = j+1$$





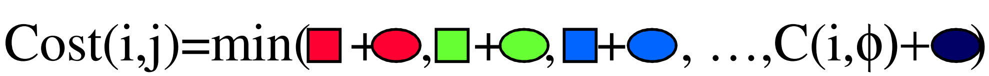
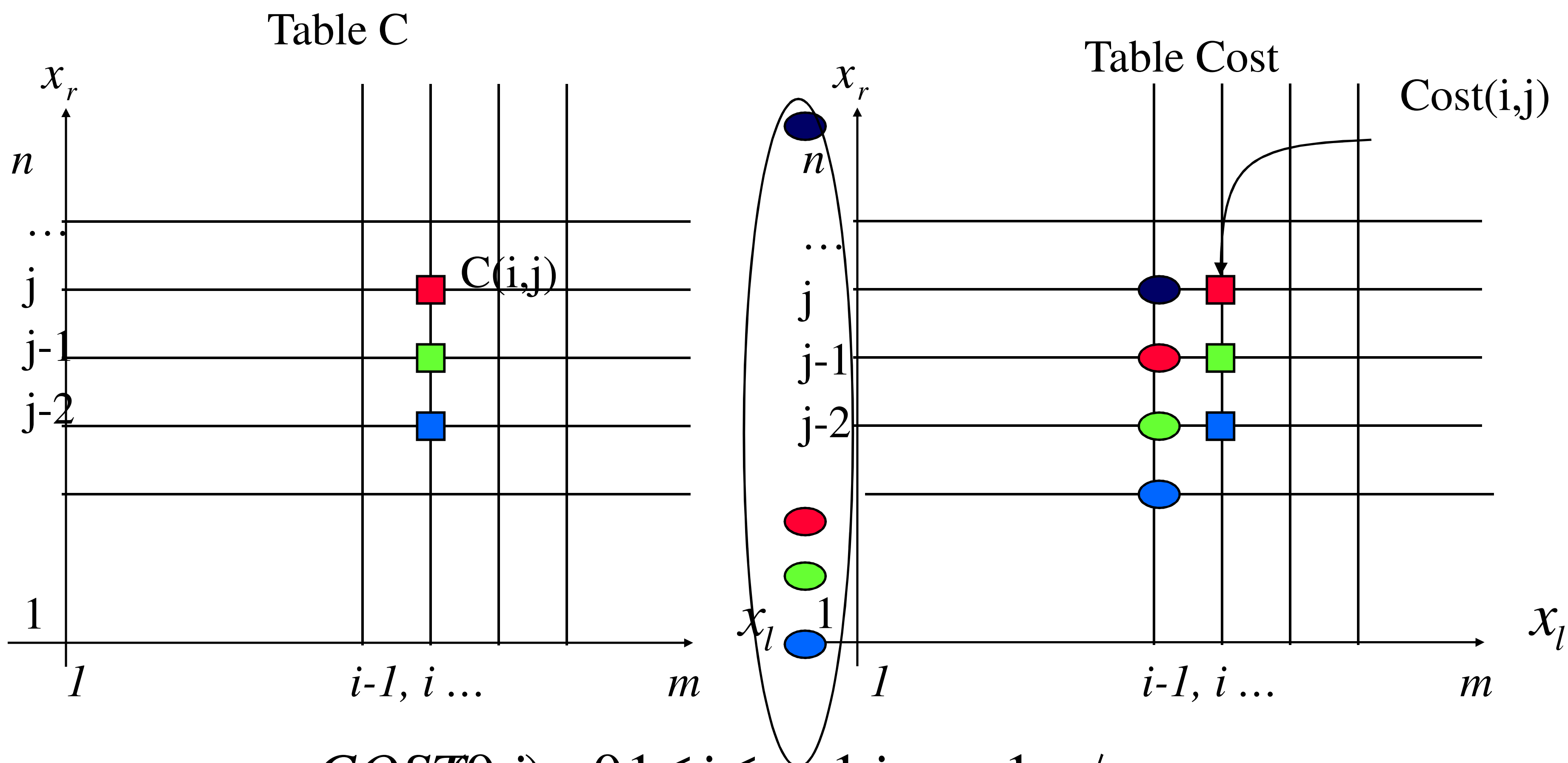


Table Building – Initial Condition

$$COST(m,n) = \min_{1\le j\le n+1}[C(m,i)+COST(m-1,i-1)], \phi = n+1$$

$$COST(i,j) = \min_{1\le k\le j+1}[C(i,k)+COST(i-1,k-1)], \phi = j+1$$



$$COST(0,i) = 0, 1\le i\le n+1, i = n+1 = \phi$$

Cost Structure

- *Cost(i,j)* has several components
  - *Self*: how compatible is the match with each other
  - *Local neighbors*: how compatible is the match in the neighborhood
    - Check changes in *d*
    - Changes are allowed if strong gradient cut through the scanline in the neighborhood
  - *Global structure*: how good is the path
    - How many matched pairs are in the path

General Local Cost Function

- Not all images are made of black dots with no apparent structures
- In fact, most images have well defined structures
- Again, large (complex) structures are more unique but w. a large range of disparity values
- Small (simple) structures are less unique but w. a small range of disparity values

Feature-Based Matching

- Edge matching
  - Filter left and right images with Gaussian of different widths
  - Edge detection
  - Match edges based on *orientation* and *strength* at coarse layers (w. fewer edges, can afford to search over a large disparity range)
  - Refine disparity at finer layers (limited disparity search range based on matches at coarser layers)

C(i,j): Local Region correlation

- Select a small window in one image
- Move the small window on the epipolar line of the other image
- Compute "similarity" (e.g., correlation coefficients)

$$\frac{\sum_i \sum_j (I^l(i,j) - \bar{I}^l(i,j))(I^r(i,j) - \bar{I}^r(i,j))}{\sqrt{\sum_i \sum_j (I^l(i,j) - \bar{I}^l(i,j))^2 \sum_i \sum_j (I^r(i,j) - \bar{I}^r(i,j))^2}}$$

Region correlation for Multi-spectral Images

$$\frac{\sum_i \sum_j (\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j))^T (\mathbf{C}^r(i,j) - \overline{\mathbf{C}}^r(i,j))}{\sqrt{\sum_i \sum_j \left\| \mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j) \right\|^2 \sum_i \sum_j \left\| \mathbf{C}^r(i,j) - \overline{\mathbf{C}}^r(i,j) \right\|^2}}$$

$$\mathbf{C} = \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

Region correlation for Multi-spectral Images with Uneven Spread – Correlation Coefficient

$$\mathbf{\Sigma} = \sum_i \sum_j (\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j))(\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j))^T$$

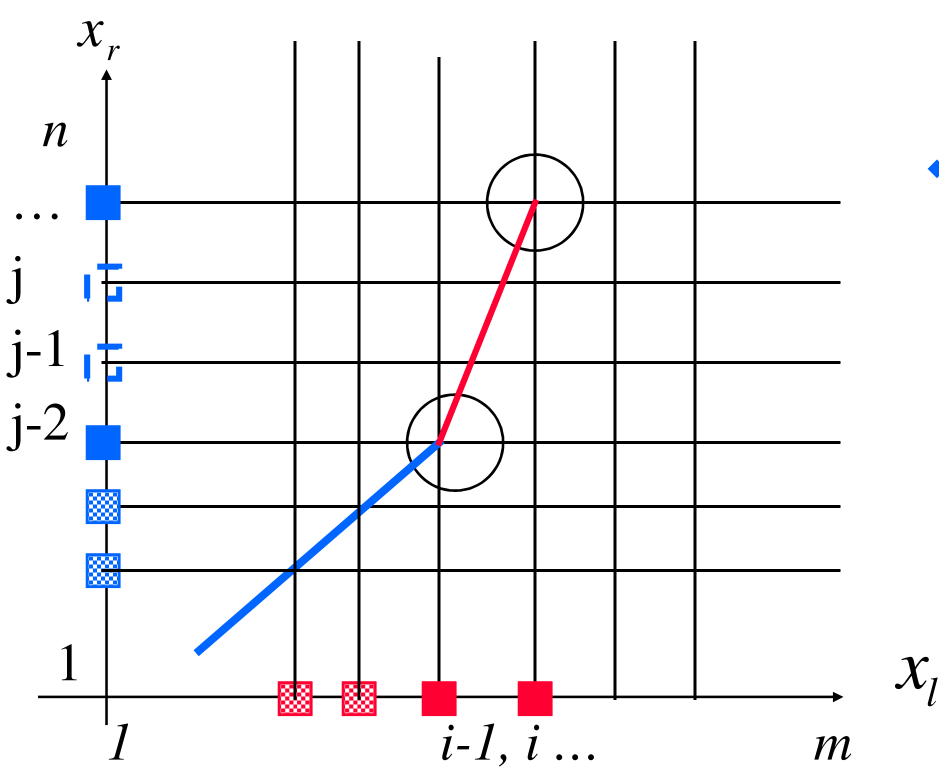
$$\frac{\sum_i \sum_j (\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j))^T \mathbf{\Sigma}^{-1} (\mathbf{C}^r(i,j) - \overline{\mathbf{C}}^r(i,j))}{\sqrt{\sum_i \sum_j [\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j)]^T \mathbf{\Sigma}^{-1} [\mathbf{C}^l(i,j) - \overline{\mathbf{C}}^l(i,j)] \sum_i \sum_j [\mathbf{C}^r(i,j) - \overline{\mathbf{C}}^r(i,j)]^T \mathbf{\Sigma}^{-1} [\mathbf{C}^r(i,j) - \overline{\mathbf{C}}^r(i,j)]}}$$

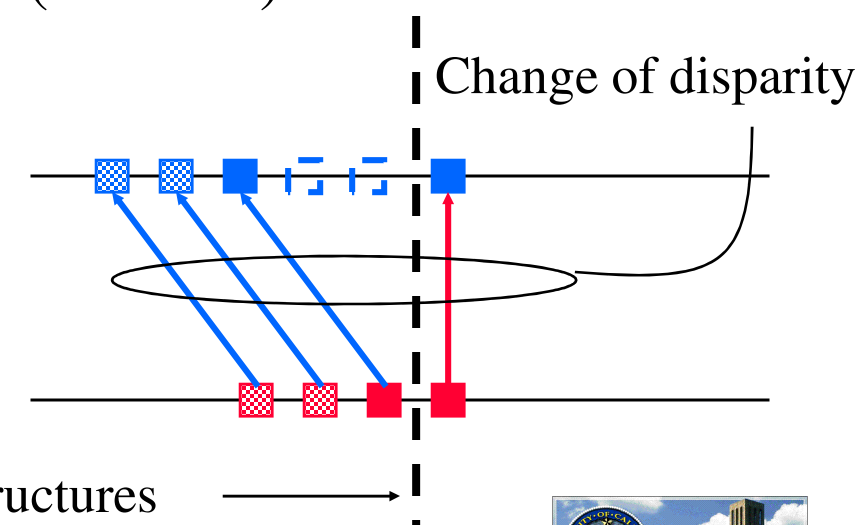
$$\mathbf{C} = \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

Local Neighborhood Cost

- "Similar" disparity can mean
  - $d=x_l-x_r$=constant , 45° lines
  - $d=x_l-x_r$=ax+b
  - Or low-order poly expression
- Change of d should be zero ($d=const$) or constant ($d=ax+b$)

Table Cost



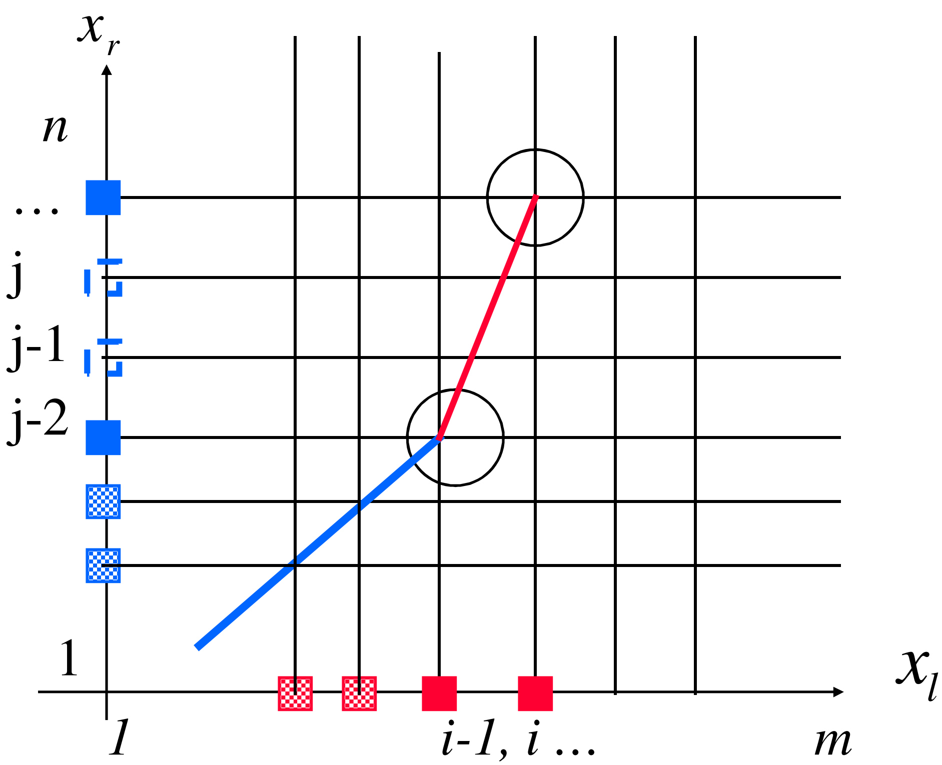


Different 3D structures

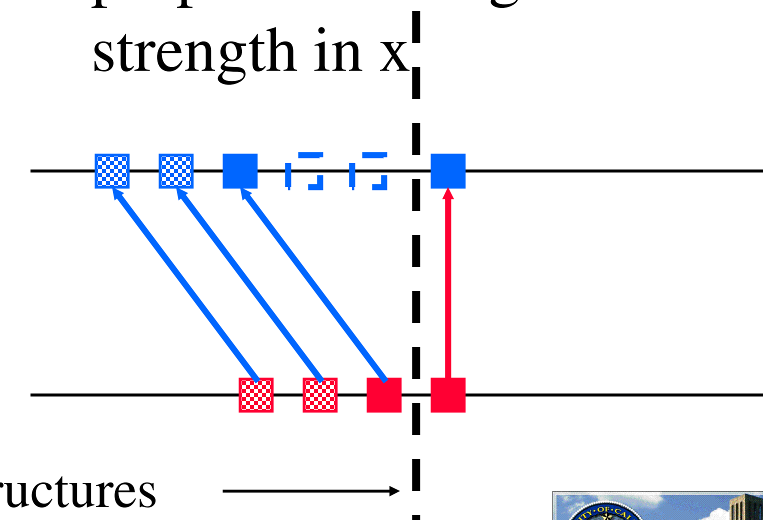
Edges should be present

Local Neighborhood

Table Cost



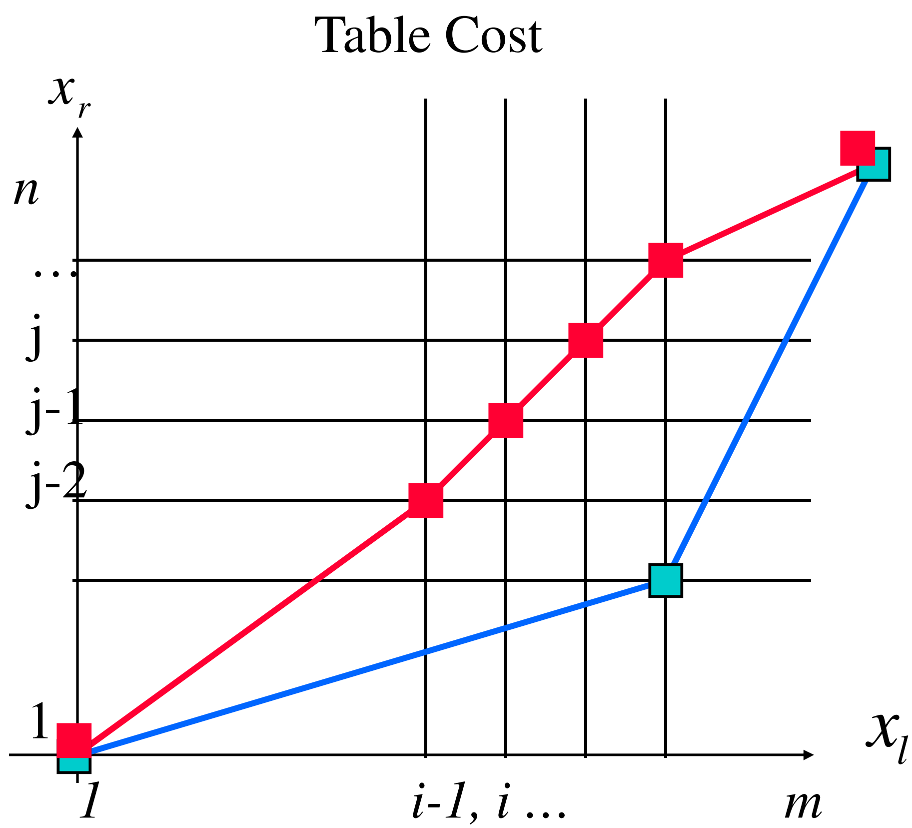
- Check disparity change in a local (one-sided) neighborhood
- If change is not zero or constant, then penalize such changes inversely proportional to gradient strength in x



Different 3D structures

Edges should be present

Global Structure



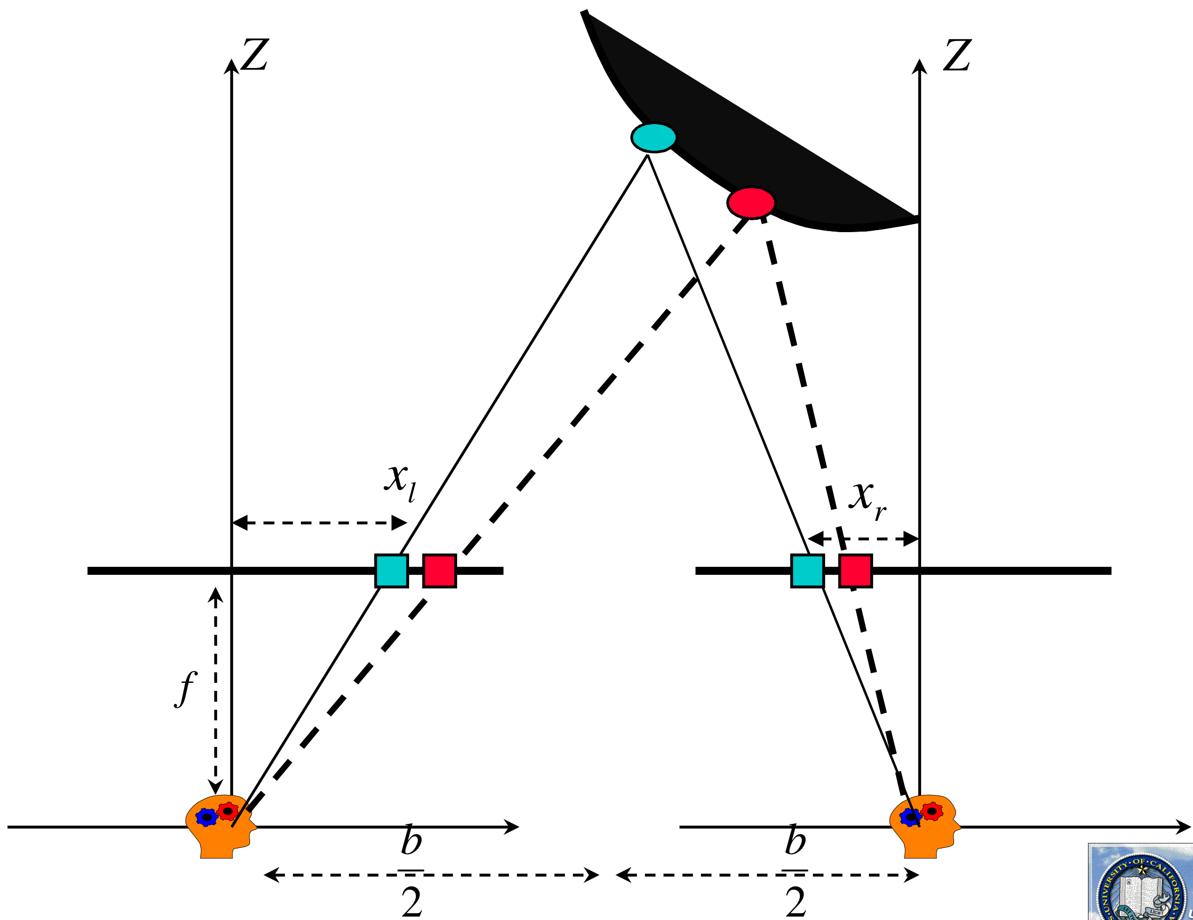
- Stereo algorithm should hopefully produce a large number of pixel matches (occasional skipping on a discrete grid is unavoidable)
- Blue curve is bad because only very few pixels are matched
- # of matched pixels should be considered

Image rectification

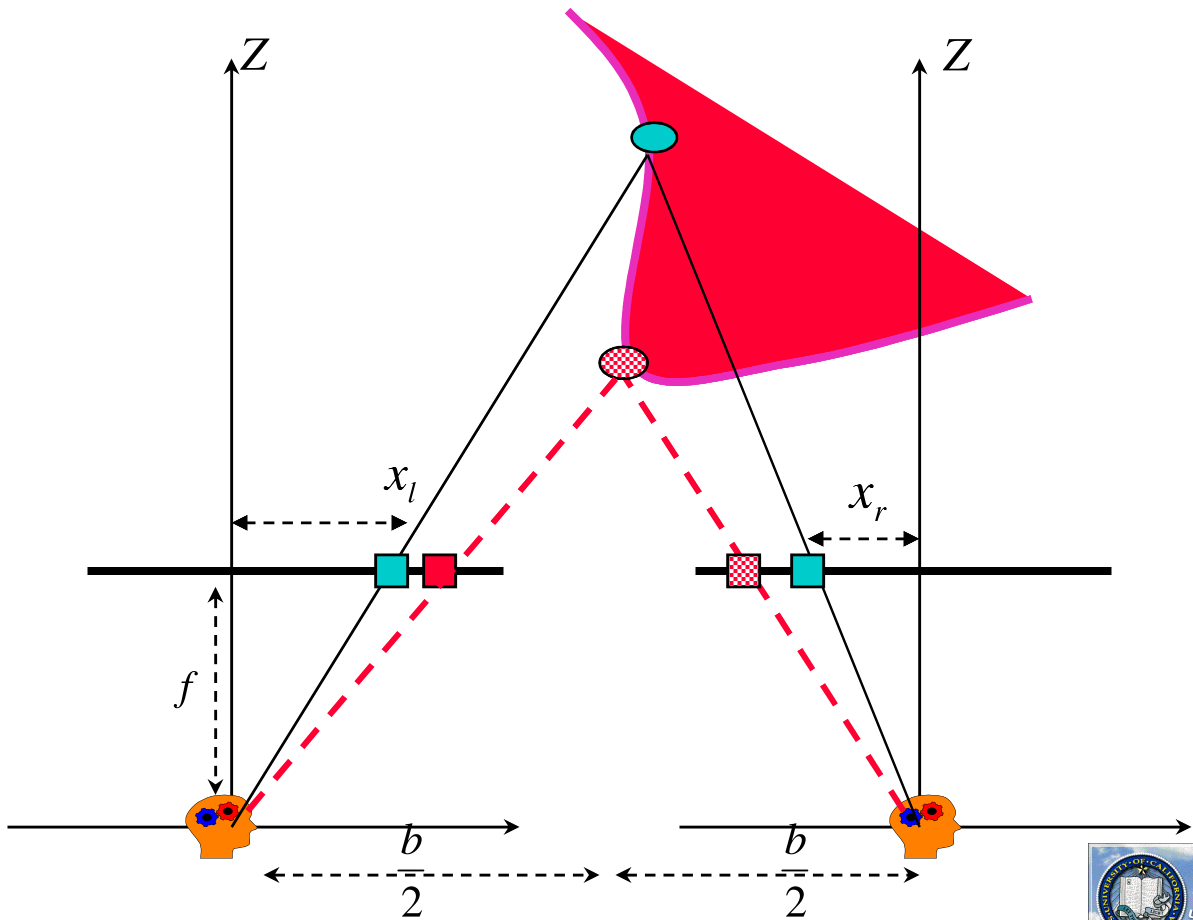
- Stereo calculations can be much simplified if the two images are *rectified* – replaced by two equivalent images with a common image plane parallel to the baseline

- Single, common image plane
- Epipolar lines are image scan lines

DP Path Constraints

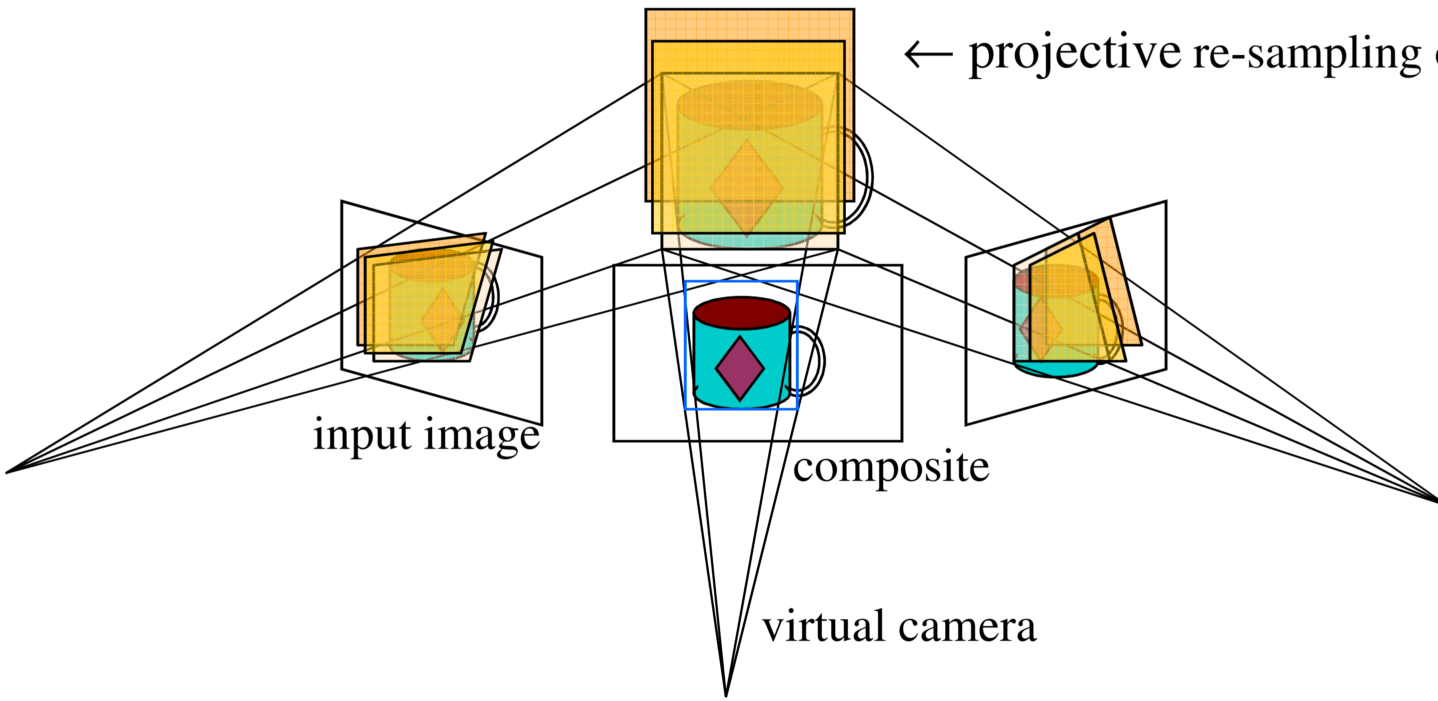


DP Path Constraints



Plane Sweep Stereo

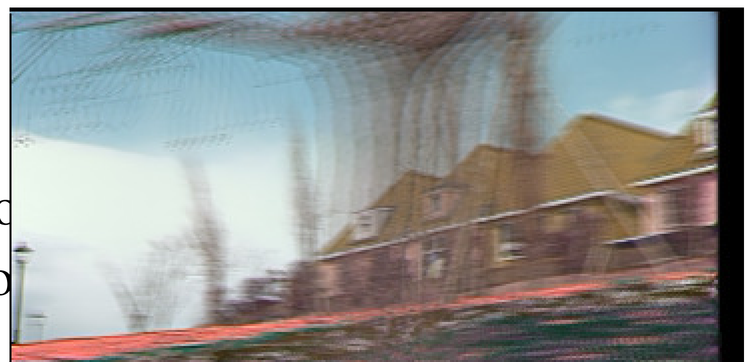
- Sweep family of planes through volume



- each plane defines an image ⇒ composite homography

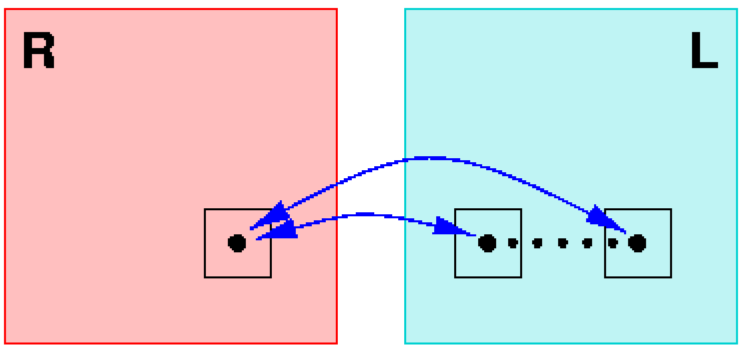
Plane Sweep Stereo

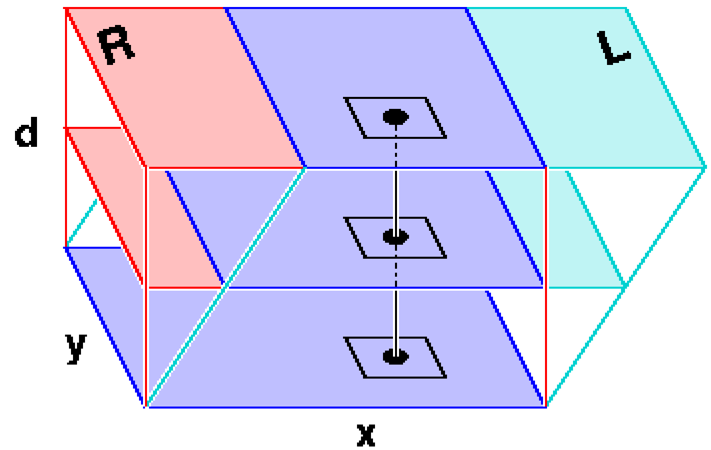
- For each depth plane
  - compute composite (mosaic) image — *mean*



  - compute erro
  - convert to co
- Select winning depth at each pixel

Plane sweep stereo

- Re-order (pixel / disparity) evaluation loops
- 



for every pixel,
  for every disparity
    compute cost

for every disparity
  for every pixel
compute cost

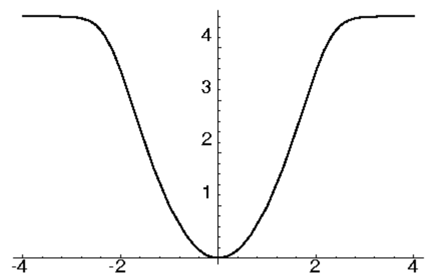
framework

1. For every disparity, compute *raw* matching costs

   Why use a robust function?

   - occlusions, other outliers

$$E_0(x, y; d) = \rho(I_L(x' + d, y') - I_R(x', y'))$$



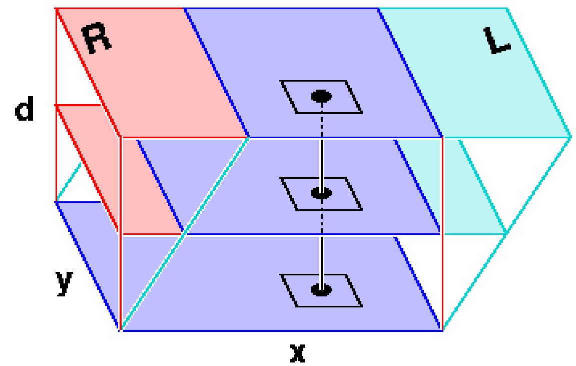
- Can also use alternative match criteria

framework

2. Aggregate costs spatially

$$E(x, y; d) = \sum_{(x',y') \in N(x,y)} E_0(x', y', d)$$

- Here, we are using a *box filter* (efficient moving average implementation)
- Can also use weighted average, [non-linear] diffusion…

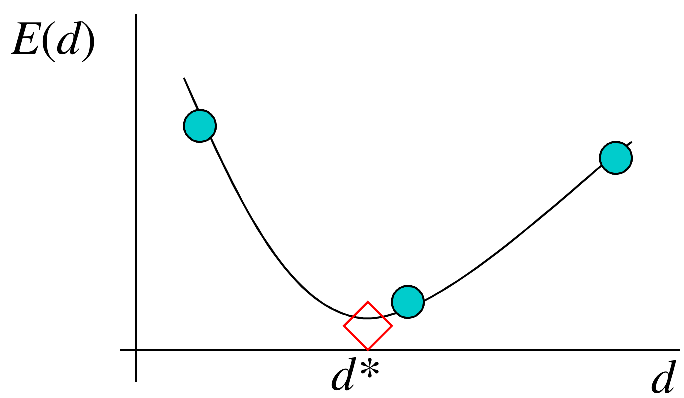


framework

3. Choose winning disparity at each pixel

4. Interpola

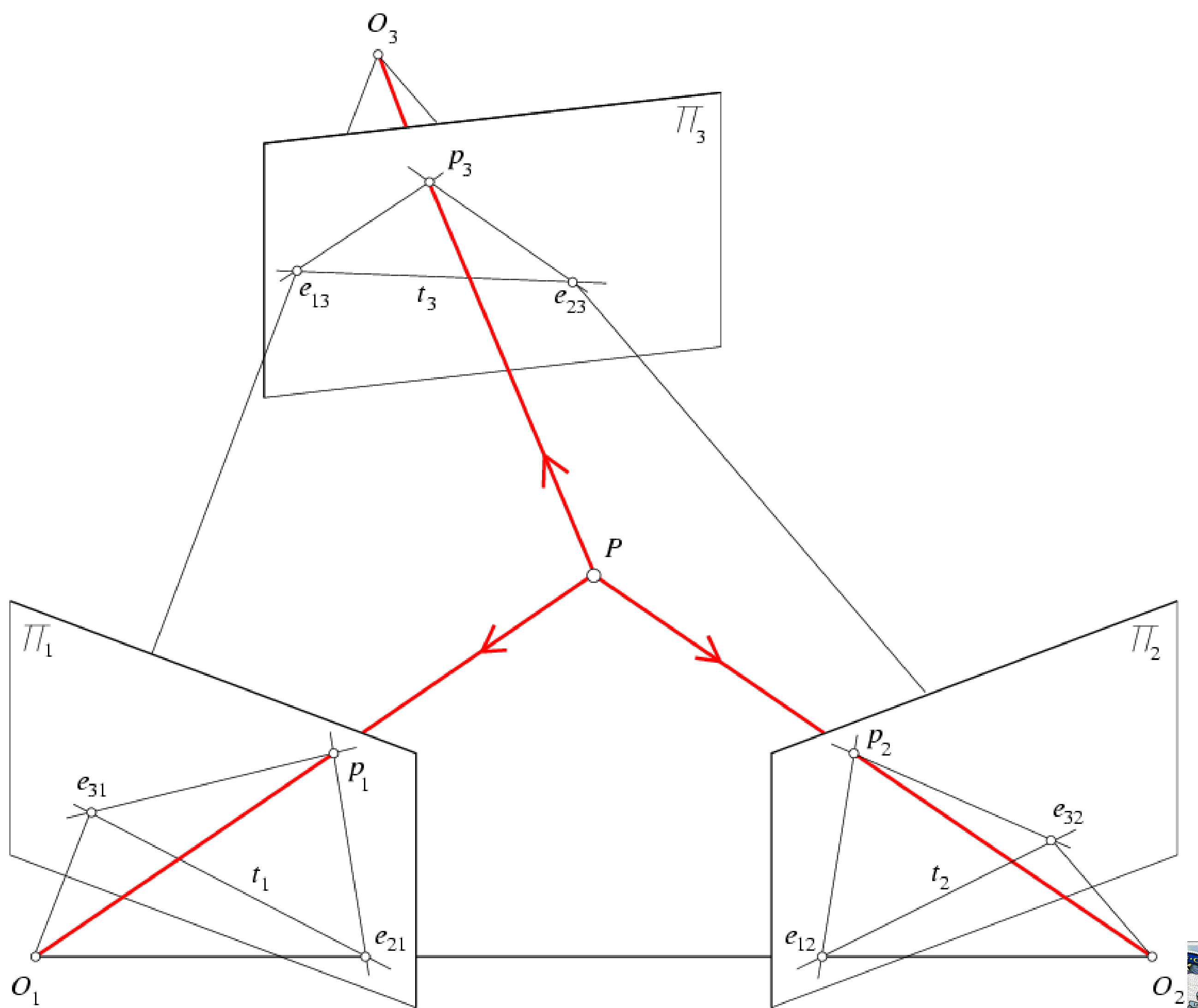
$$d(x, y) = \arg\min_d E(x, y; d)$$



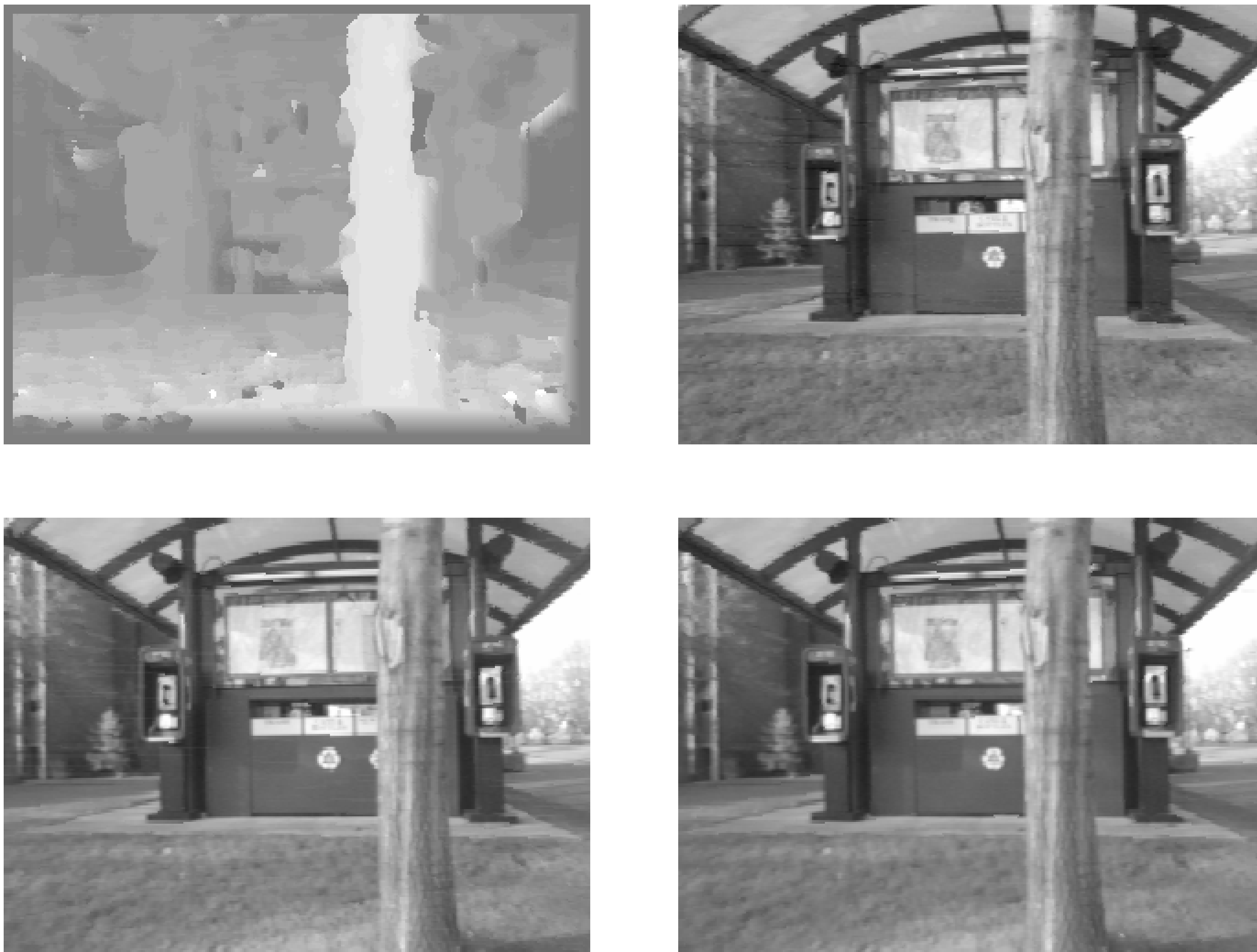
Multiple camera stereo

- Using multiple camera in stereo has advantages and disadvantages
- Some disadvantages
  - Computationally more expensive
  - More correspondence matching issues
  - More hardware ($)
- Some advantages
  - Extra view(s) reduces ambiguity in matching
  - Wider range of view, fewer “holes”
  - Better noise properties
  - Increased depth precision

Trinocular (three-view) epipolar constraint



3-camera stereo

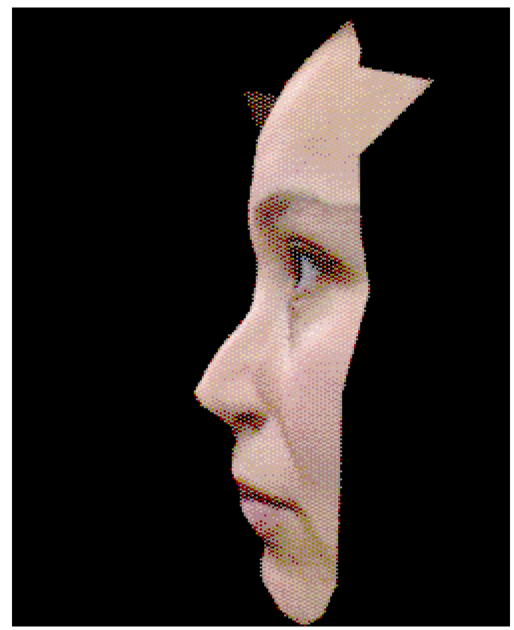


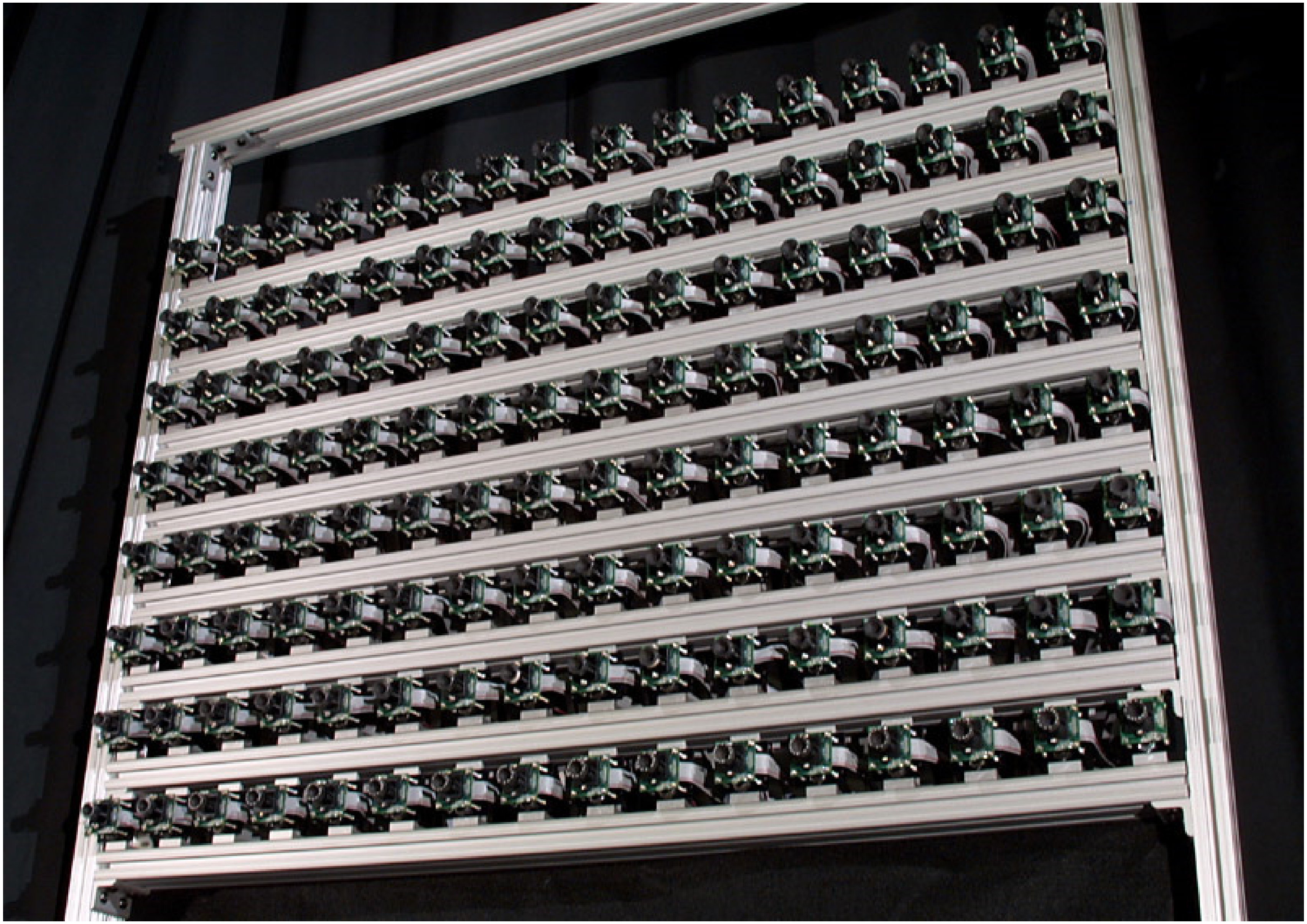
Example: Four views

Univ. of Penn

Input images

Texture input

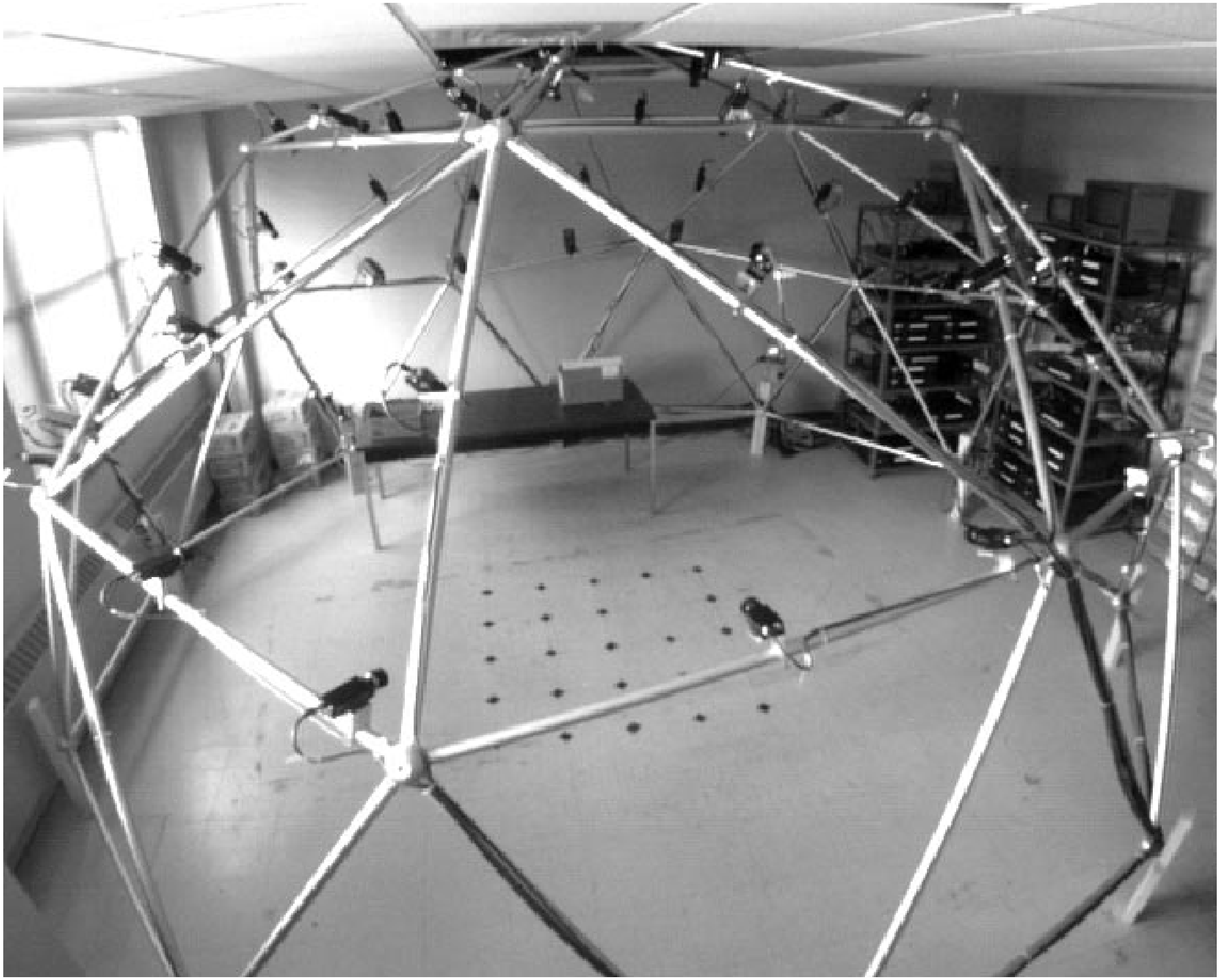




The Stanford Multi-Camera Array

128 CMOS cameras, 2” baseline

5x5 racks version: 125 CMOS cameras, 9” baseline
4 capture PCs, 4 electronics racks (1 board per camera)

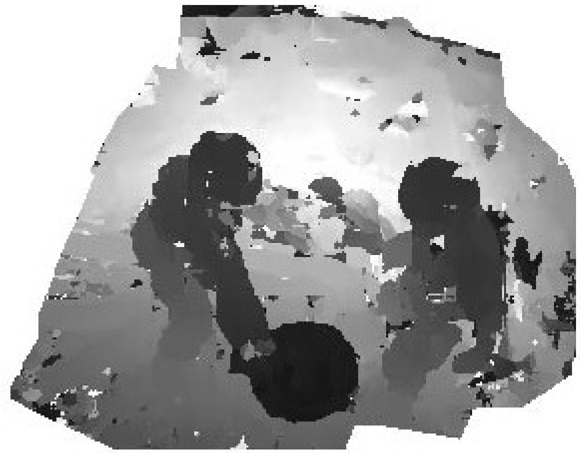


CMU multi-camera stereo

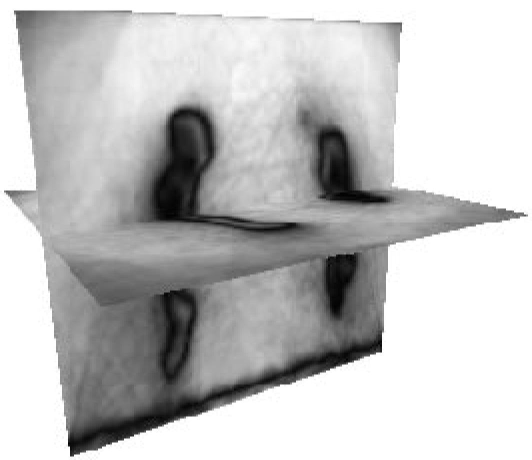
51 video cameras mounted on a 5-meter diameter geodesic dome.

Example: Basketball

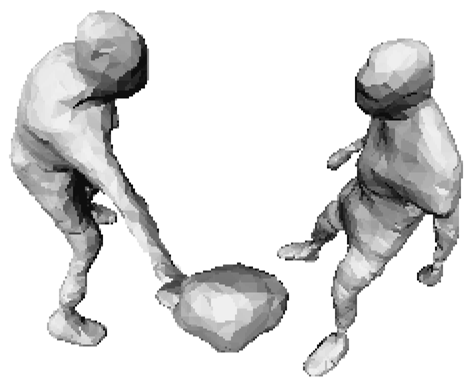
(a)



(b)



(c)



(d)

a) Original scene

b) Range Image

c) Integrated range images

d) 3D model extraction

Example: Basketball (cont.)

(e)

(f)

e) Rendered view of model with texture

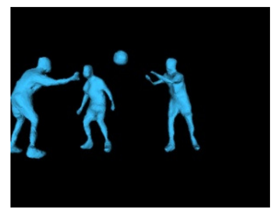
f) Rendered view of model from a virtual camera and combined with another digitized scene

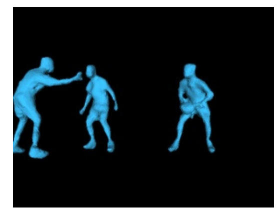
Example: Basketball (cont.)

Inputs (two separate events)

Video 1

Reconstructed 3D shape



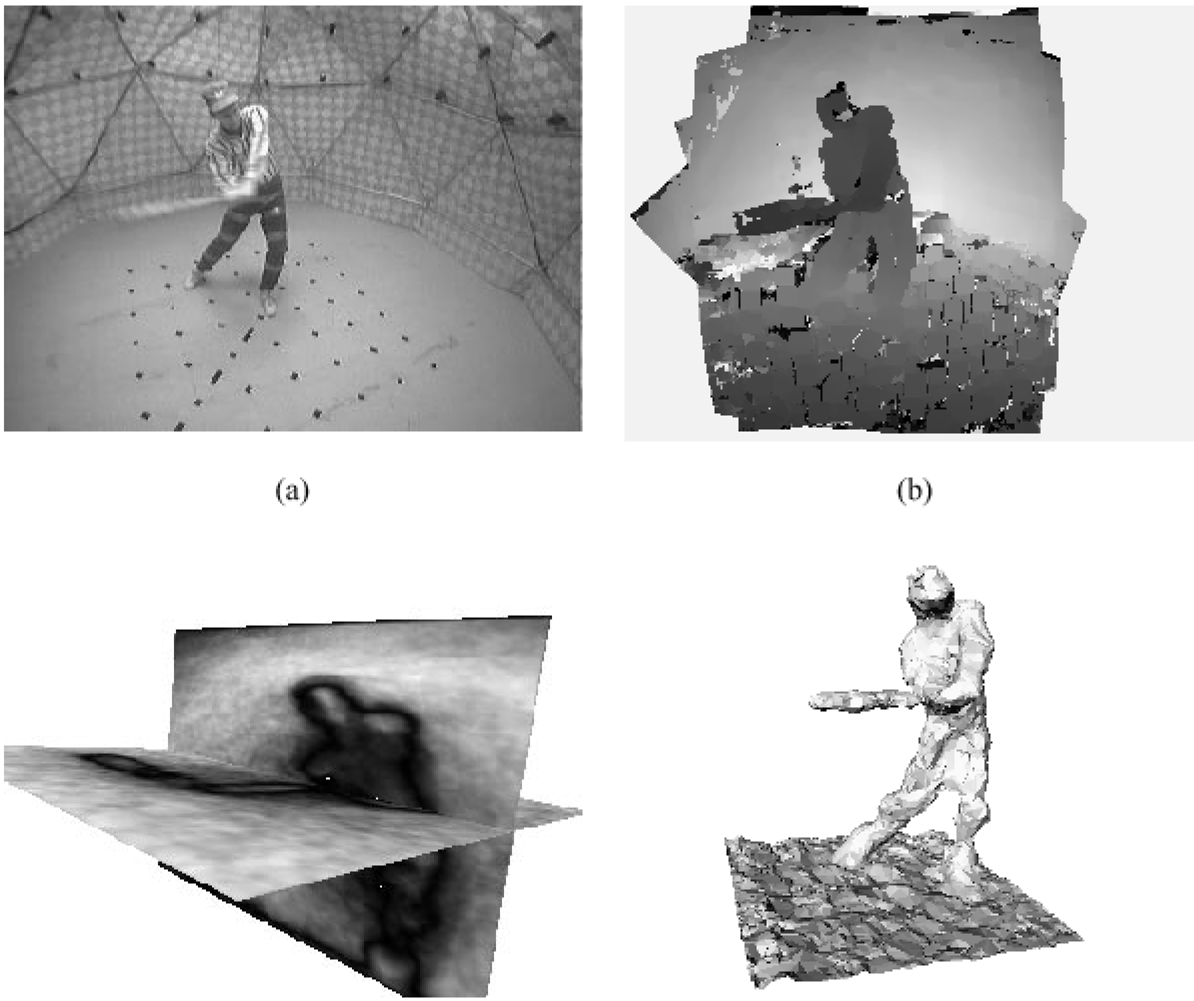


Video 2

Virtual View of combined event

Video 3

Example: Baseball



(a) (b) (c) (d)

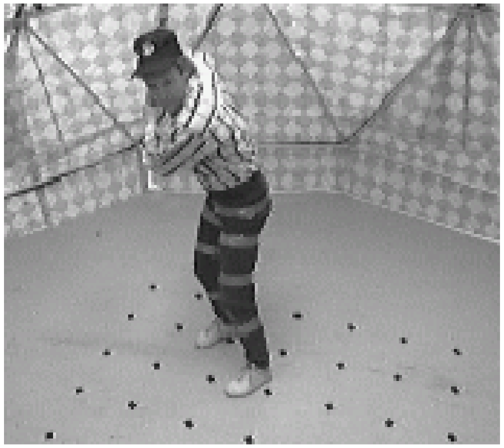
a) Original scene

b) Range Image

c) Integrated range images

d) 3D model extraction

Example: Baseball (cont.)



This example features a person swinging a baseball bat inside the recording studio. A director might select a single camera that provides a good view of the swing from the side (as in the above), but you might prefer to

- circle around as the batter swings...
- or stop the batter
- drop from above...
- *be the BALL!*