LECTURE 08: STEREO

Venu Madhav Govindu
Department of Electrical Engineering
Indian Institute of Science, Bengaluru

2023

Stereo

- In this lecture we shall consider image **stereopsis**
- Basic geometry of stereo images
- Estimation Considerations
- Structured-Light Stereo and Depth Cameras

Stereo

Stereo images

- Reconstruct scene depth given two images
- Images are taken from two different viewpoints
- The difference in images “encodes” depth information
- Classic problem in computer and human vision (binocular stereopsis)
- *stereo* Greek root meaning “solid” (PIE root *ster meaning “stiff”)
- Great advances in both
  - Understanding image geometry
  - Algorithmic solution to stereopsis
- We shall consider only the basic issues underlying stereo
- We shall consider a simple formulation and generalise later

Parallax

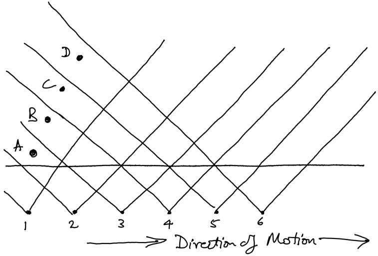
- What do you observe ?
- Different *apparent* relative speeds
- Why ?
- Role of depth

Parallax

- What do you observe ?
- Different *apparent* relative speeds
- Why ?
- Role of depth

Parallax

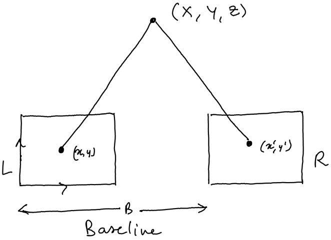
- What do you observe ?
- Different *apparent* relative speeds
- Why ?
- Role of depth



Parallax

- What do you observe ?
- Different *apparent* relative speeds
- Why ?
- Role of depth

Stereo

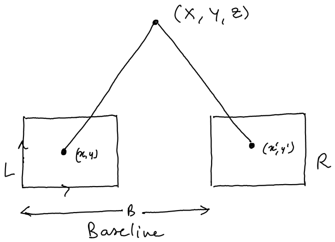


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x,y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x,y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

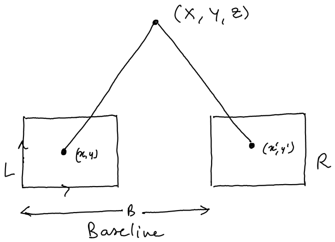


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x,y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x,y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

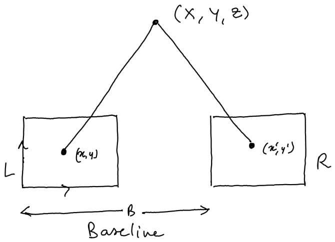


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x, y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x, y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

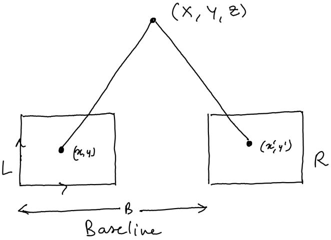


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x, y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x, y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

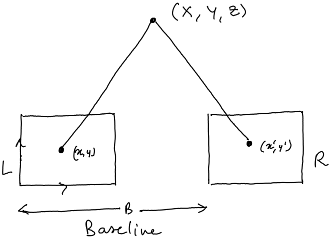


- Left image projection ?
  - $x = \frac{fX}{Z}$; $y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}$; $y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}$; $y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x,y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x,y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

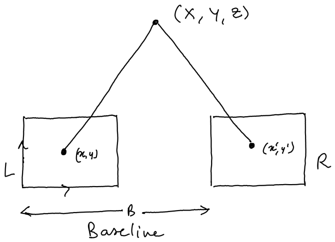


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $d(x, y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}$
- What does $d(x, y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

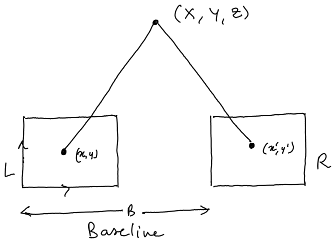


- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x,y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x,y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo



- Left image projection ?
  - $x = \frac{fX}{Z}; y = \frac{fY}{Z}$
- Right image?
  - $x' = \frac{f(X+B)}{Z}; y' = \frac{fY}{Z}$
- Relations between images ?
  - $x' - x = \frac{fB}{Z}; y' = y$

Canonical Stereo

- Pure horizontal translation (eyes)
- Horizontal shift of cameras (B is baseline)
- Disparity $\boxed{d(x,y) = x - x' = \frac{fB}{Z} \propto \frac{1}{Z}}$
- What does $d(x,y)$ mean ?
- Distant vs. near objects ?
- Vertical shift ?

Stereo

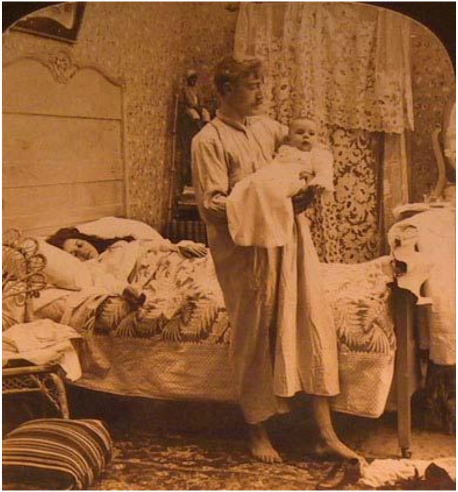
Depth illusion achieved by cross-fusion

Picture taken from camelphotos.com

Stereo

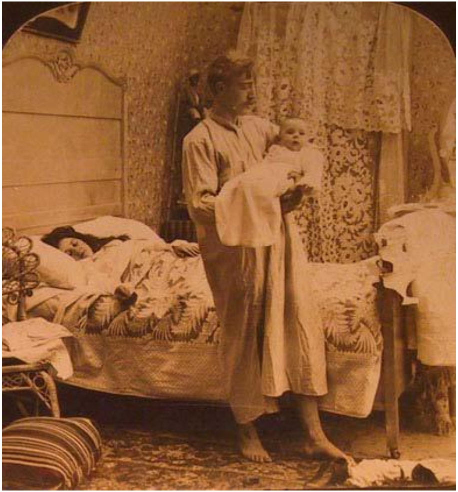
picture taken from www.antiquemystique.com

Stereo



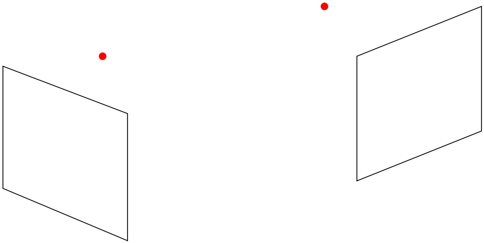
picture taken from slides of Michael Black

Stereo



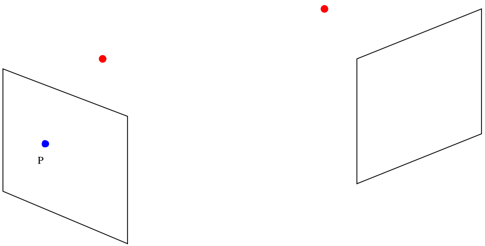
picture taken from slides of Michael Black

Stereo



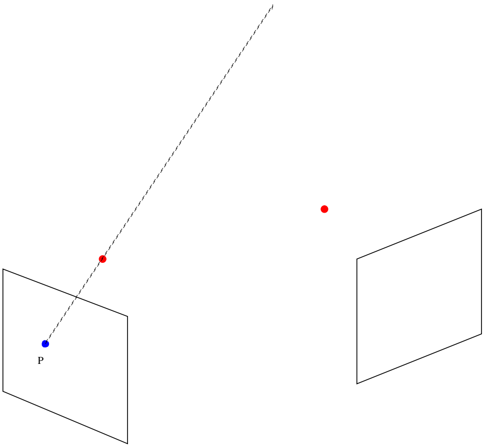
Consider two cameras looking at a scene

Stereo



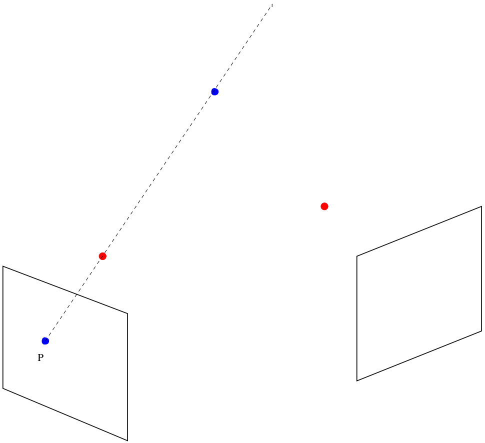
"Inverse" ray given an image point

Stereo



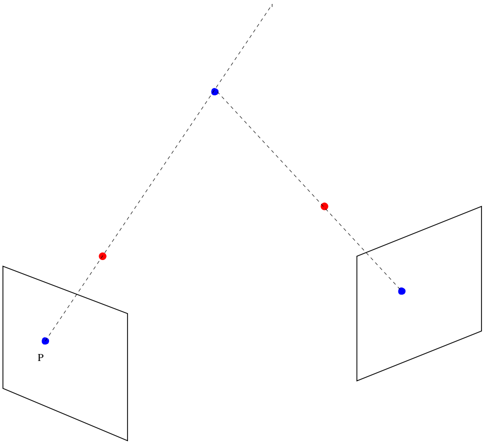
Putative 3D point

Stereo



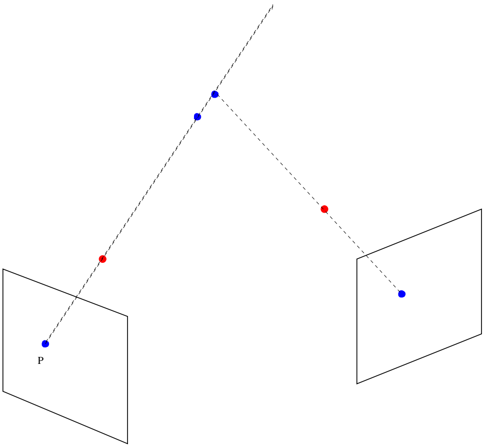
Putative point projects into second image

Stereo



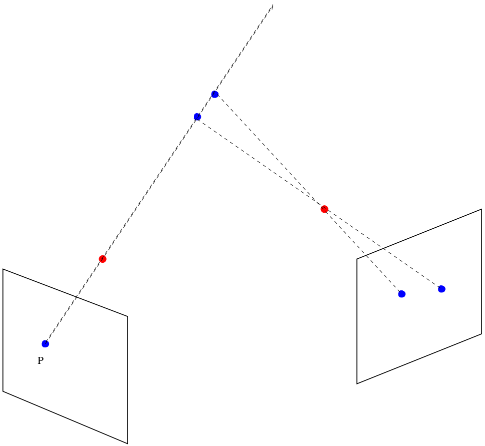
What if its somewhere else ?

Stereo



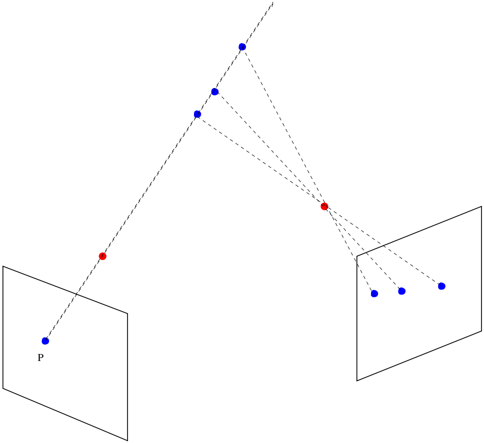
What if its somewhere else ?

Stereo



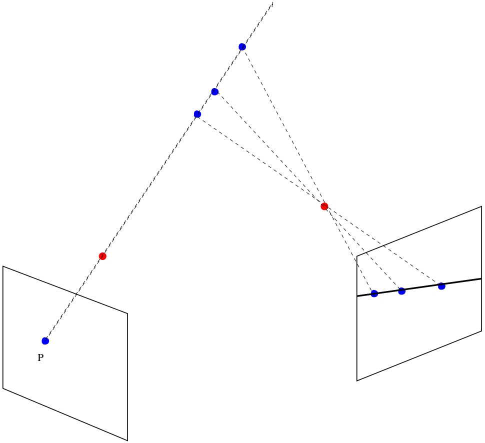
What if its somewhere else ?

Stereo

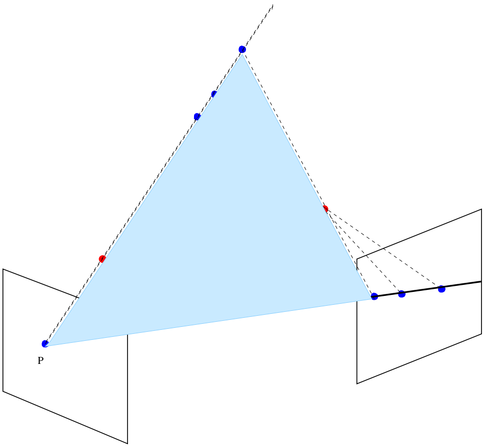


What if its somewhere else ?

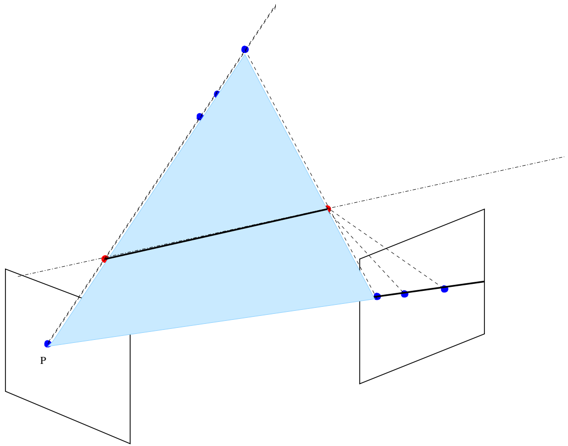
Stereo



Possible 3D points project to a line

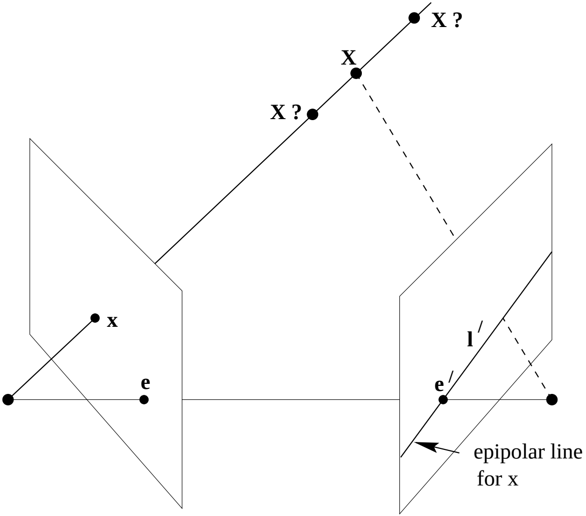


All of these sit on a plane

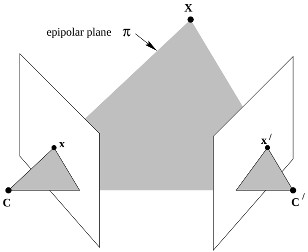


Line joining optical centers common to all planes

Stereo



Stereo

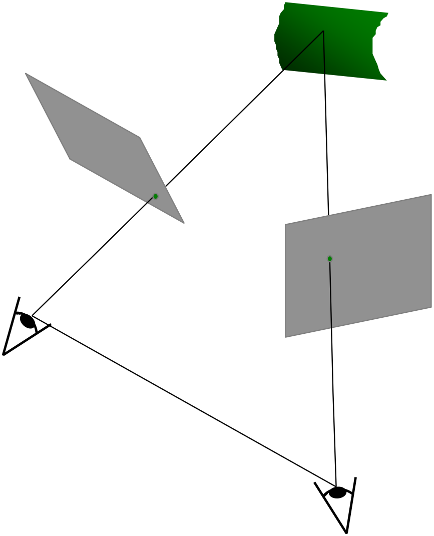


Epipolar Plane

- A **key** insight in multi-view geometry
- Can generate a succint summary of camera geometry
- Leads to a nice formulation
- What happens when cameras have pure horizontal translation?
- Can solve for camera geometry given enough point matches
- Actual algorithm will be considered in later lectures
- For now we shall focus on generic "stereo" framework

Following 3 slides borrowed from Alyosha Efros lectures

Stereo image rectification



Stereo image rectification

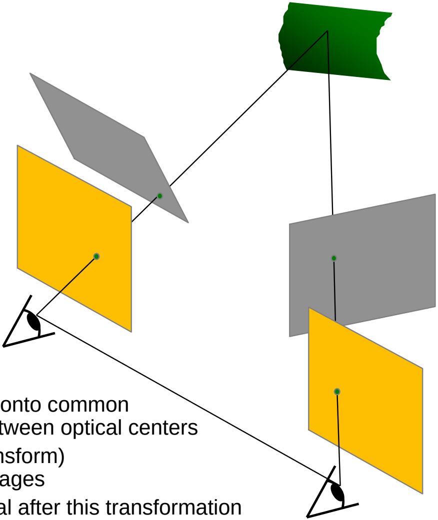
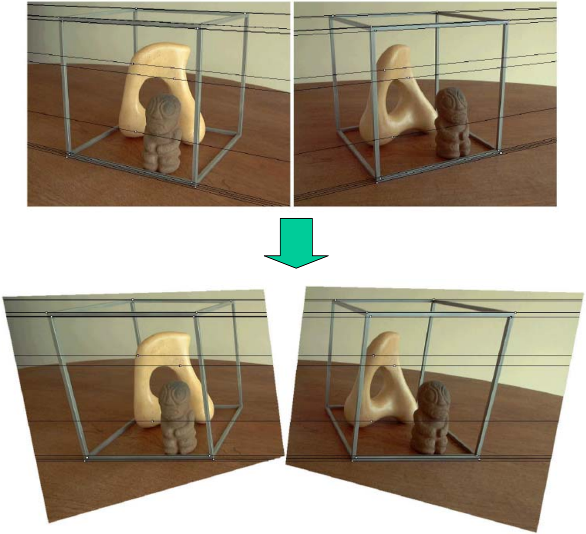


Image Reprojection

- reproject image planes onto common plane parallel to line between optical centers
- a homography (3x3 transform) applied to both input images
- pixel motion is horizontal after this transformation
- C. Loop and Z. Zhang. [Computing Rectifying Homographies for Stereo Vision](). IEEE Conf. Computer Vision and Pattern Recognition, 1999.

Stereo Rectification



Stereo

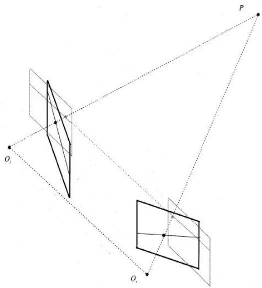
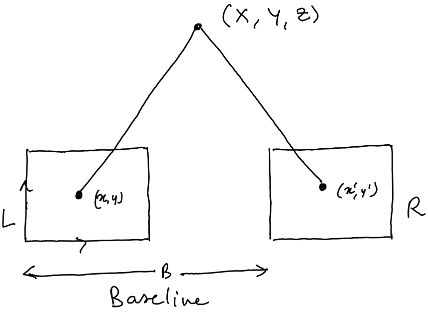


Figure 7.8 Rectification of a stereo pair. The epipolar lines associated to a 3-D point P in the original cameras (black lines) become collinear in the rectified cameras (light grey). Notice that the original cameras can be in any position, and the optical axes may not intersect.

Rectification

- Rotate image planes
- Send epipoles to infinity
- Scale appropriately
- Will not consider details here

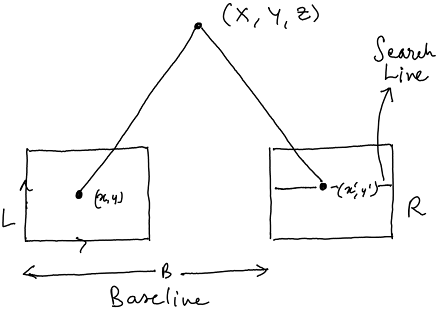
Stereo



Issues in Canonical Stereo

- Geometry is simple and fixed
- Correspondence $\Longleftrightarrow$ Disparity $\Longleftrightarrow$ Depth
- How to get correspondences ?
- Other refinements are crucial

Stereo



Search Space?

- We only need to search along a line for a match
- Greatly reduces search space
- Want general line along row of pixels

Stereo

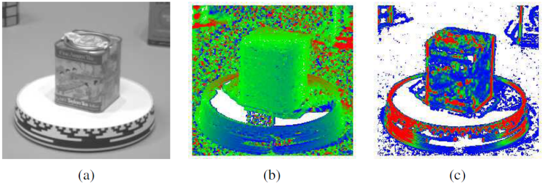


Figure 12.11 *Uncertainty in stereo depth estimation (Szeliski 1991b): (a) input image; (b) estimated depth map (blue is closer); (c) estimated confidence(red is higher). As you can see, more textured areas have higher confidence.*

Issues to be addressed

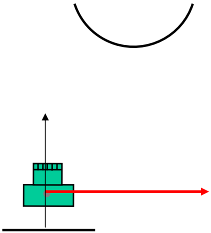
- What to match ? (Features, Patches)
- How to look for a match ? (Search strategy)
- What constraints can we enforce ?

What to match ?

- Brightness values or intensities
- Points (corners)
- Edges
- **Patches** Why this ?

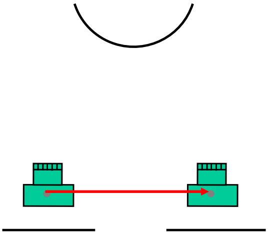
Following slides are borrowed from Michael Black lectures

Binocular Stereo



Left

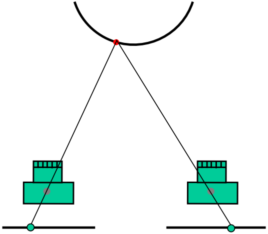
Binocular Stereo



Left

Right

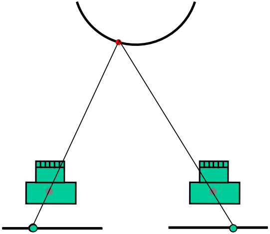
Binocular Stereo



Left

Right

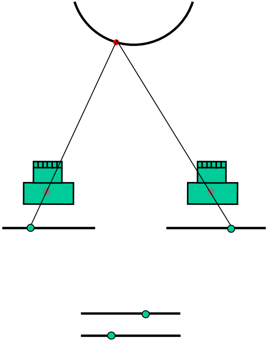
Binocular Stereo



Left

Right

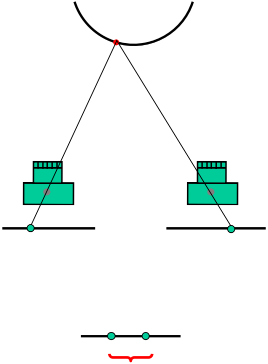
Binocular Stereo



Left

Right

Binocular Stereo

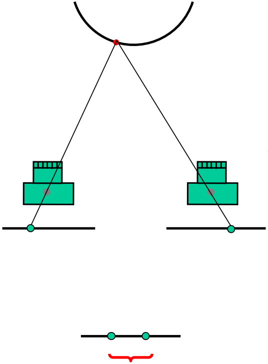


Left

Right

binocular disparity

Binocular Stereo



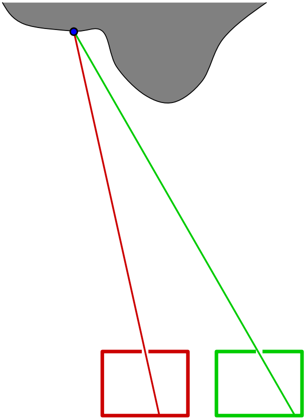
From known geometry of the cameras and estimated disparity, recover depth in the scene

Left

binocular disparity

Right

Stereo Geometry

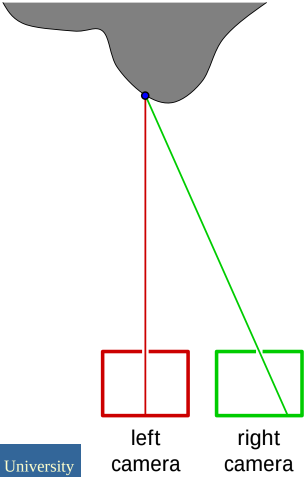


left camera

right camera

Scharstein

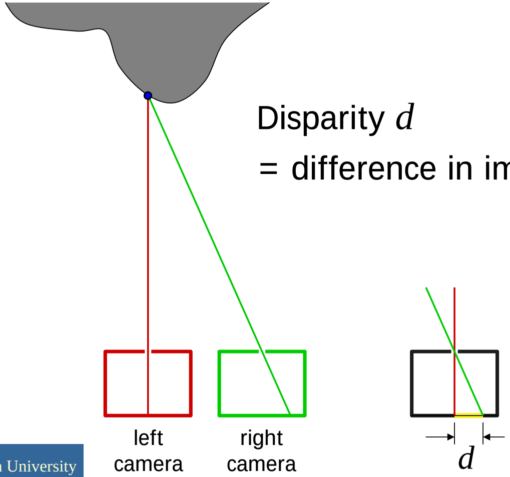
Stereo Geometry



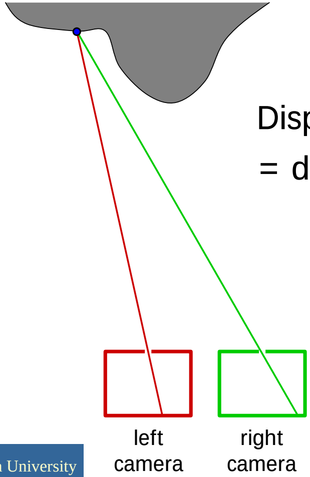
Stereo Geometry

Disparity d

= difference in image position

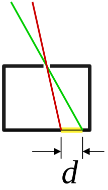


Stereo Geometry

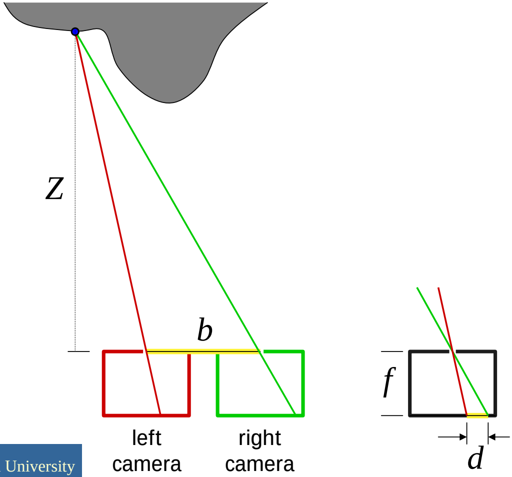


Disparity d

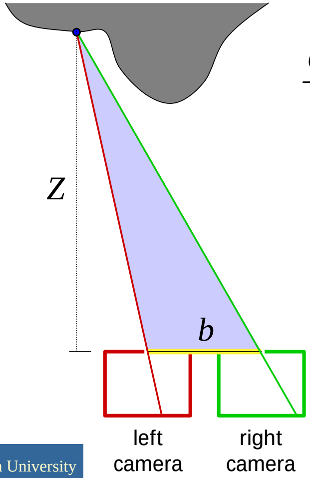
= difference in image position



Stereo Geometry

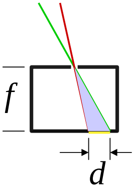


Stereo Geometry



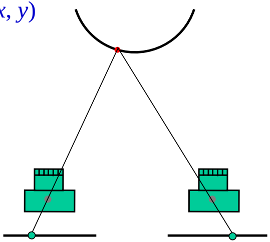
$$\frac{d}{f} = \frac{b}{Z}$$

Disparity $d = bf\,\frac{1}{Z}$



Binocular Disparity

$Z(x, y)$ is depth at pixel (x, y)
$d(x, y)$ is disparity



Estimate:

$$Z(x,y) = \frac{f\,B}{d(x,y)}$$

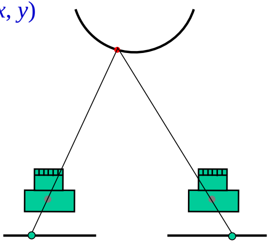
Left

Right

Search for best match

Binocular Disparity

$Z(x, y)$ is depth at pixel (x, y)
$d(x, y)$ is disparity

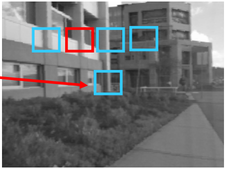


Estimate:

$$Z(x,y) = \frac{f\,B}{d(x,y)}$$

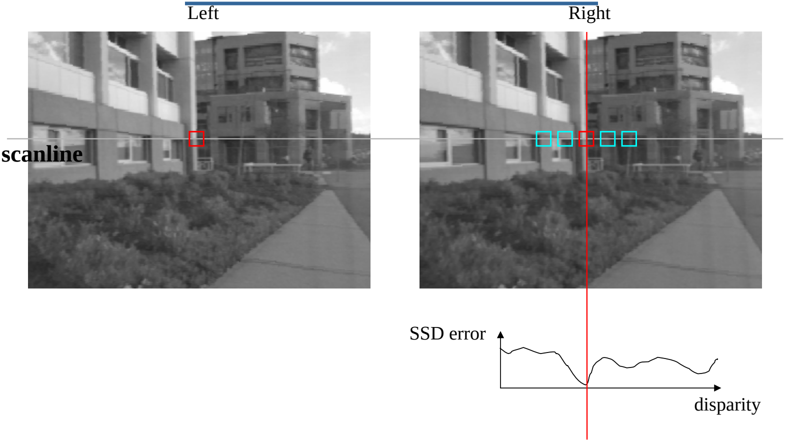
Left

Right

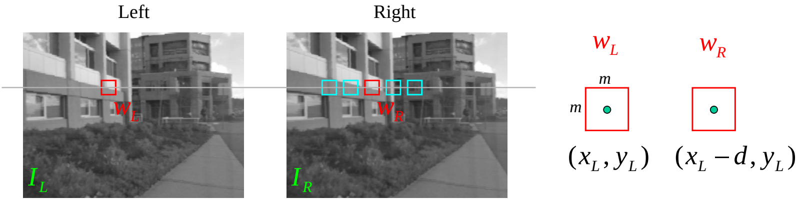


Do I need to consider this region?

Correspondence Using Correlation



Sum of Squared (Pixel) Differences



w_L and w_R are corresponding m by m windows of pixels.

The SSD cost measures the intensity difference as a function of disparity :

$$SSD_r(x, y, d) = \sum_{(x',y') \in W_m(x,y)} (I_L(x', y') - I_R(x' - d, y'))^2$$

Matching

- Even when the cameras are identical models, there can be differences in gain and sensitivity.
- The cameras do not see exactly the same surfaces, so their overall light levels can differ.
  - occlusion

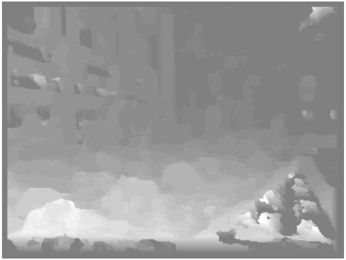
$$E_r(x,y,d) = \sum_{(x',y')\in W_m(x,y)} \rho(I_L(x',y') - I_R(x'-d,y'))$$

Robust matching function.

Correspondence Using SSD

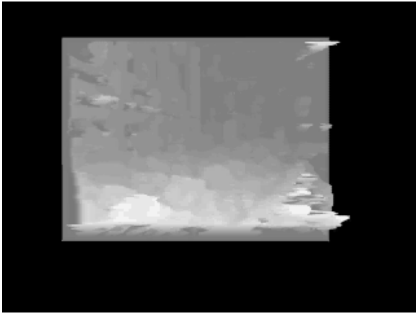
Left

Disparity Map



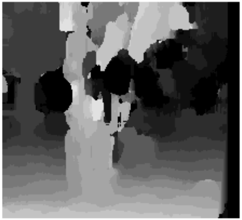
Images courtesy of Point Grey Research

Stereo Results



Window size

W = 3



W = 20

Better results with *adaptive window*

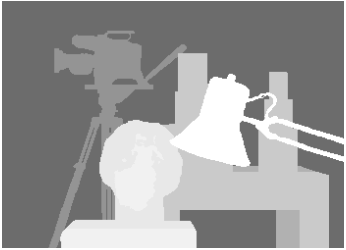
- T. Kanade and M. Okutomi, A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment,, Proc. International Conference on Robotics and Automation, 1991.
- D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. International Journal of Computer Vision, 28(2):155-174,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

(Seitz)

Stereo results

– Data from University of Tsukuba

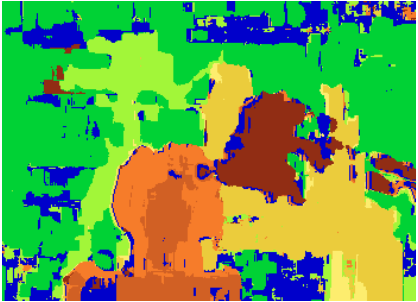
Scene



Ground truth

(Seitz)

Results with window correlation



Window-based matching
(best window size)

Ground truth

(Seitz)

Results with better method

State of the art method

Ground truth

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

(Seitz)

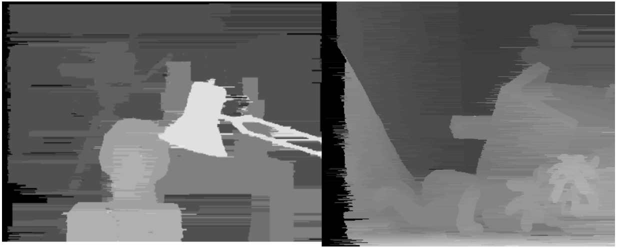
Stereo

Matching Strategies

- Brute Force search
- Coarse-to-fine search (multi-resolution pyramids)
- Relaxation
- Dynamic Programming
- MRF models

Adapted from slides of Chuck Dyer

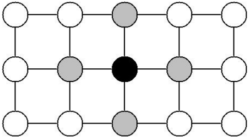
Stereo



Smoothness and Robustness

- Smoothness of disparity. Why?
- Piecewise smooth model
- Dynamic Programming for scanline
- Why do we have streaky depths?
- Remedy: 2D Models

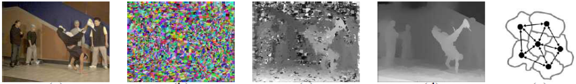
Stereo



Markov Random Fields (MRF)

- Data Term: $C(x, y, d(x, y))$ (stereo disparity cost)
- Equivalent to $P(\boldsymbol{x}|\theta)$
- Conditional probablity on neighbours $P(\theta)$
- Robust (piecewise)smoothness as a prior assumption
- Smoothness Term: $\rho(d(x, y) - d(N(x, y)))$
- Optimise: (Data Term) + λ (Smoothness Term)
- Results in joint optimisation of all disparities
- Used extensively for many problems
- Efficient discrete methods (multiway graphcuts)

CVOnline: 'Markov Random Field Optimisation' by Peter Orchard



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

Figure 12.14 *Segmentation-based stereo matching (Zitnick, Kang et al. 2004) © 2004 ACM: (a) input color image; (b) color-based segmentation; (c) initial disparity estimates; (d) final piecewise-smoothed disparities; (e) MRF neighborhood defined over the segments in the disparity space distribution (Zitnick and Kang 2007) © 2007 Springer.*

Markov Random Fields (MRF)

- Data Term: $C(x, y, d(x, y))$ (stereo disparity cost)
- Equivalent to $P(\boldsymbol{x}|\theta)$
- Conditional probablity on neighbours $P(\theta)$
- Robust (piecewise)smoothness as a prior assumption
- Smoothness Term: $\rho(d(x, y) - d(N(x, y)))$
- Optimise: (Data Term) + λ (Smoothness Term)
- Results in joint optimisation of all disparities
- Used extensively for many problems
- Efficient discrete methods (multiway graphcuts)

CVOnline: 'Markov Random Field Optimisation' by Peter Orchard

Stereo

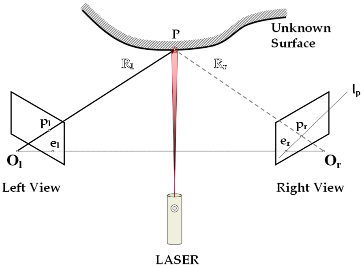
Topics not considered

- Many sophisticated optimisation methods
- Over-segmented patches and aggregation
- Multiview stereo and space carving
- Role of learning in stereo
- **Learning to match patches**
- Monocular depth estimation!

Stereo

Active Light Stereo

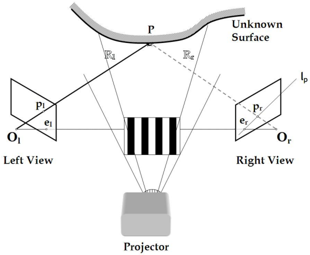
- Stereo with ambient lighting
- Recall textureless regions
- Use an active light source
- Many sensors available now
  - Structured-Light Stereo
  - Time-of-Flight Cameras
  - LIDAR (Light Detection and Ranging)



Active Light Stereo

- Ambiguous: Textureless regions
- Thought exercise: single laser spot (accurate, easy, slow)
- Project a pattern
- Projector geometry equivalent to pinhole camera
- Different ways of establishing correspondence

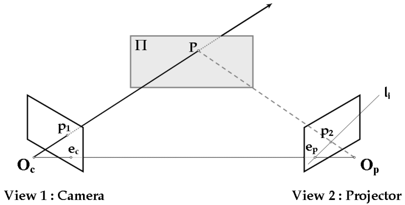
Stereo



Active Light Stereo

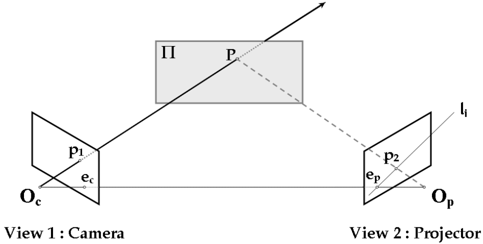
- Ambiguous: Textureless regions
- Thought exercise: single laser spot (accurate, easy, slow)
- Project a pattern
- Projector geometry equivalent to pinhole camera
- Different ways of establishing correspondence
  - Binary Encoding
  - Phase Encoding

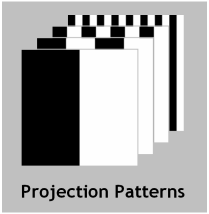
Stereo



Active Light Stereo

- Ambiguous: Textureless regions
- Thought exercise: single laser spot (accurate, easy, slow)
- Project a pattern
- Projector geometry equivalent to pinhole camera
- Different ways of establishing correspondence
    - Binary Encoding
    - Phase Encoding

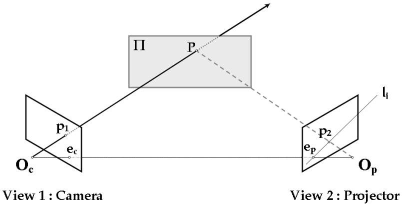




Active Light Stereo

- Ambiguous: Textureless regions
- Thought exercise: single laser spot (accurate, easy, slow)
- Project a pattern
- Projector geometry equivalent to pinhole camera
- Different ways of establishing correspondence
    - Binary Encoding
    - Phase Encoding

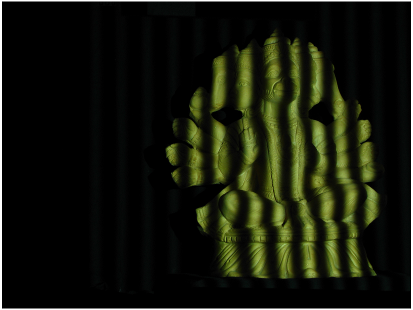
Stereo



Active Light Stereo

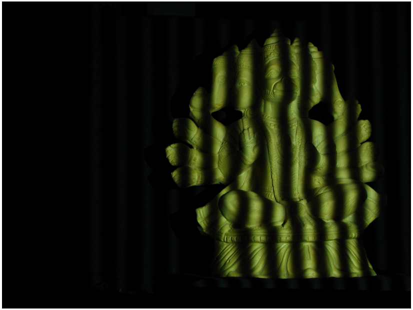
- Ambiguous: Textureless regions
- Thought exercise: single laser spot (accurate, easy, slow)
- Project a pattern
- Projector geometry equivalent to pinhole camera
- Different ways of establishing correspondence
  - Binary Encoding
  - Phase Encoding

Stereo



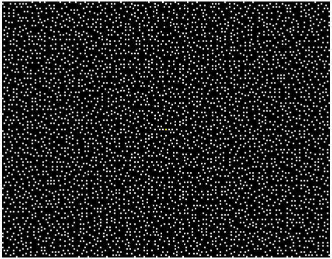
High Accuracy 3D Scan (0.1 mm)

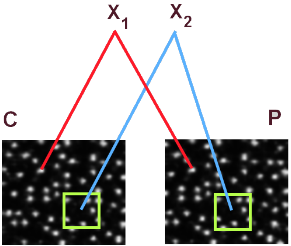
Stereo



High Accuracy 3D Scan (0.1 mm)

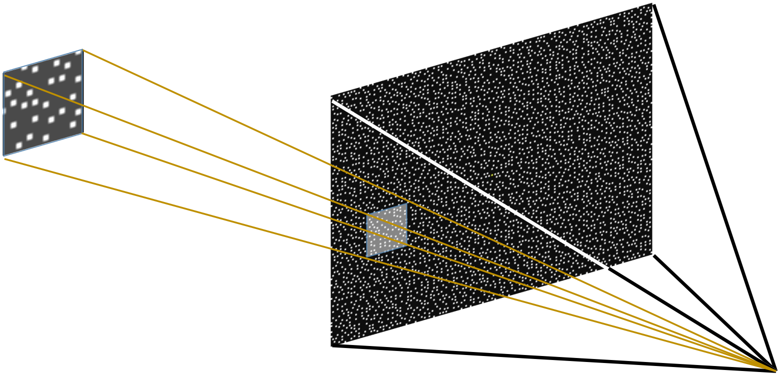
Structured-Light Stereo

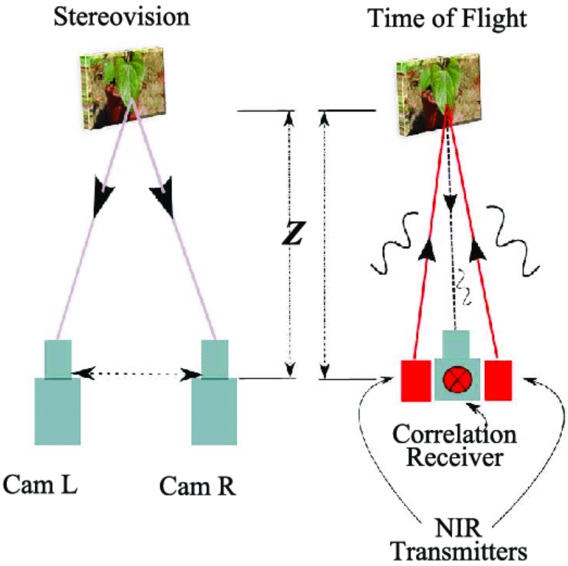




Projecting Random Dot Pattern

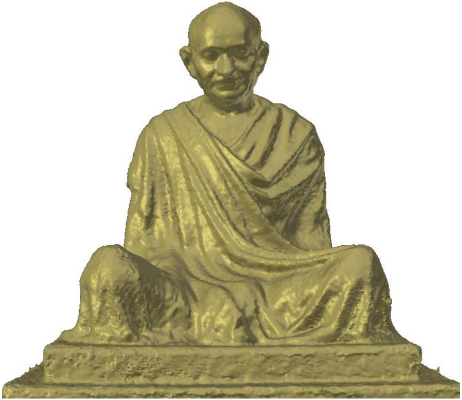
- Kinect (First Version) and other depth cameras
- Single shot scanning (low power infra-red laser)
- Ensures uniform amount of texture on surface
- Random pattern $\Rightarrow$ uniqueness of patch
- Easier to match using unique patches
- Many improvements, also ToF cameras





Wajahat Kazmi *et al.*, 'Indoor and Outdoor ...'

Stereo



Statue of Mahatma Gandhi at Sabarmati Ashram, Ahmedabad (90 cm height)