Stereo Geometry

DD2423 Image Analysis and Computer Vision

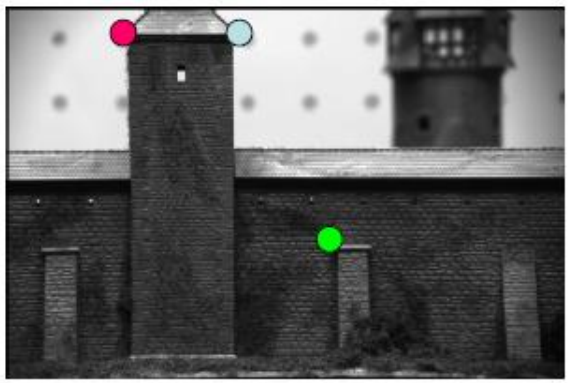
Mårten Björkman

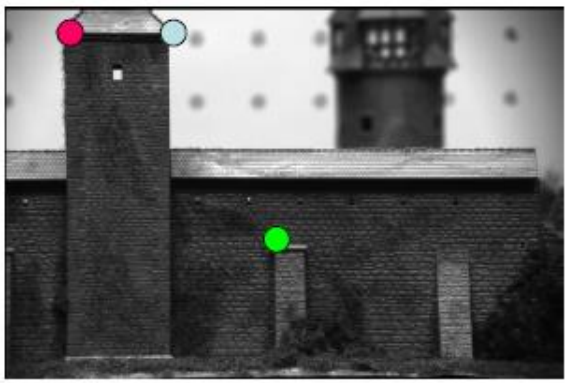
Robotics, Perception and Learning Division
School of Electrical Engineering and Computer Science

December 8, 2021

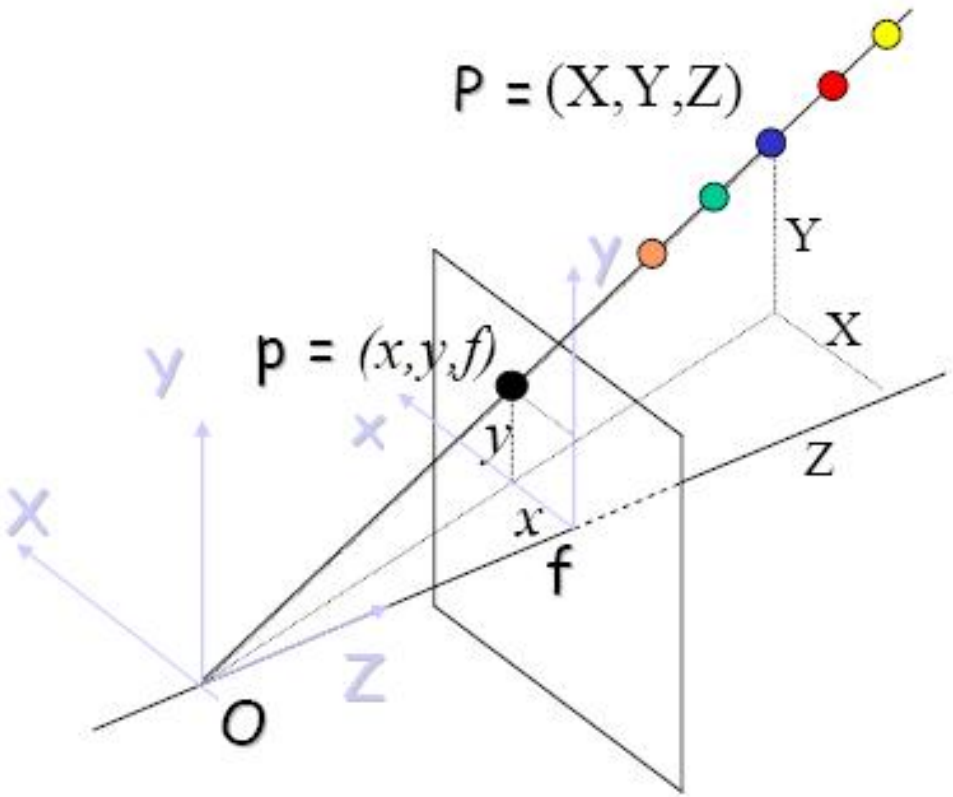
Stereo

- Inferring depth from images taken at the same time by two or more cameras by using the differences between object's positions.
- If corresponding points can be identified in the left and right images, depth can be computed by triangulation.





Why stereo?

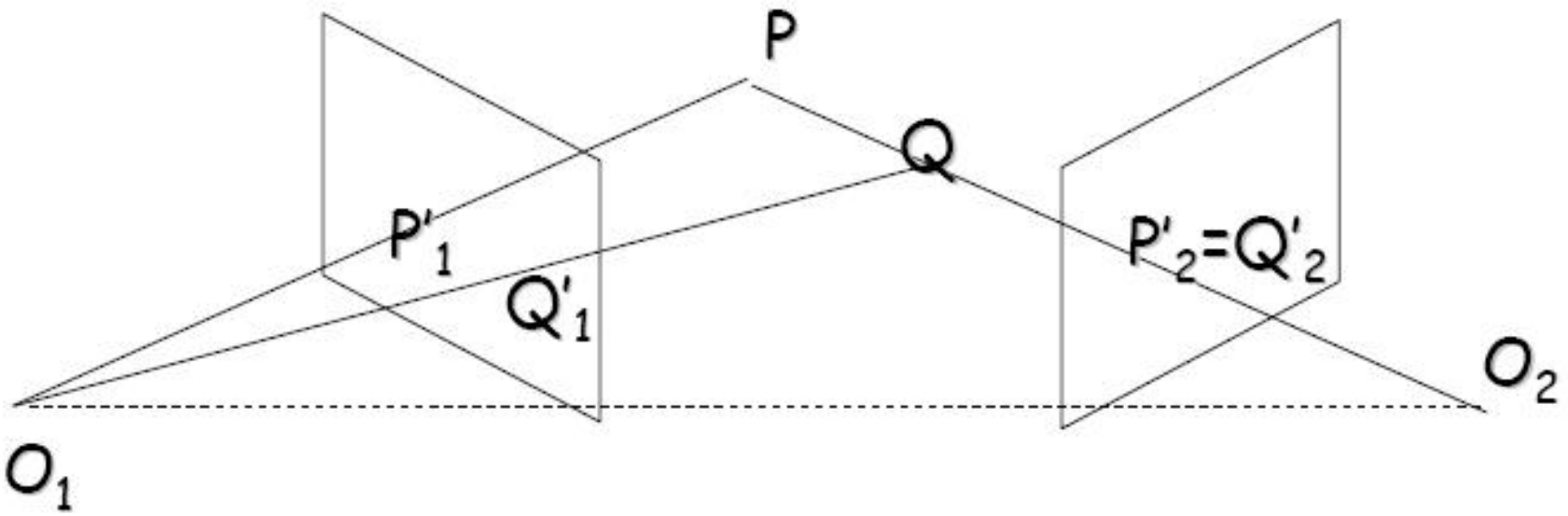


$$x = f\frac{X}{Z} = f\frac{kX}{kZ}$$

$$y = f\frac{Y}{Z} = f\frac{kY}{kZ}$$

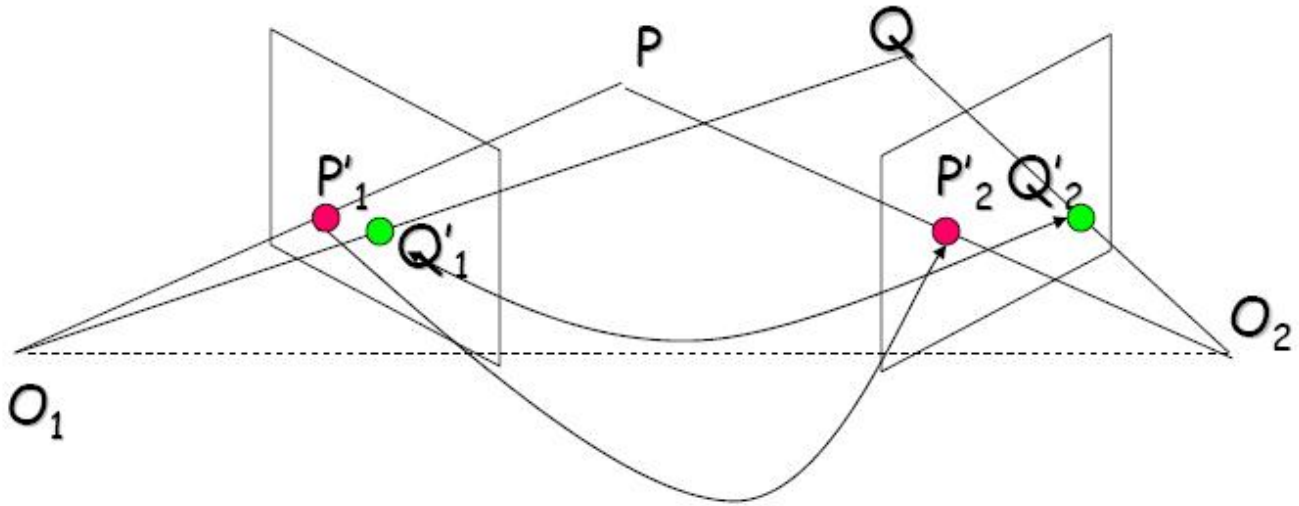
Fundamental Ambiguity:
Any point on the ray OP has image p

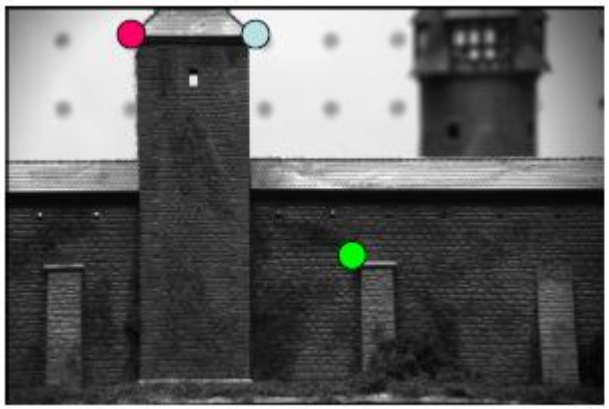
Stereo

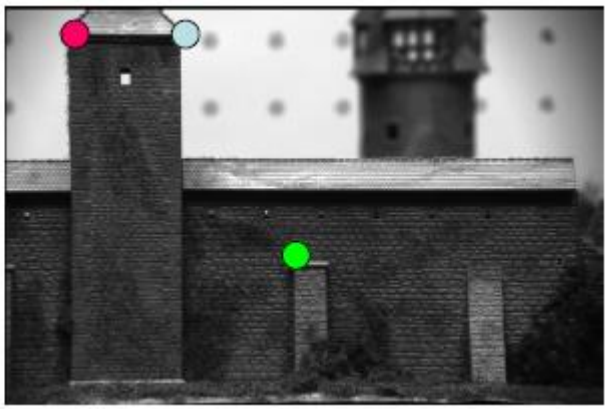


Ambiguities may still exist, for example due to occlusions.
Two 3D points gets projected to the same point in one of the images.

Stereo







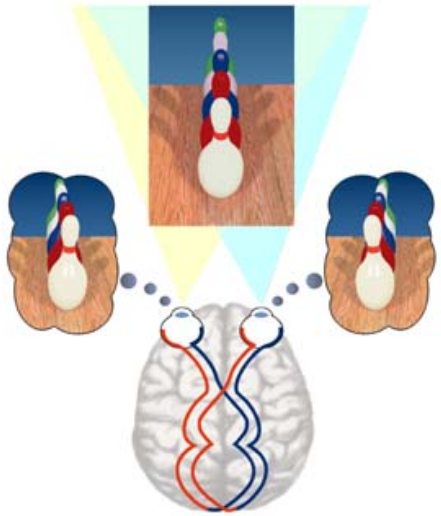
Simple test of your stereo vision

- Hold a pen in each hand, at about a relaxed arms length, place the ends toward each other separated about 10cm.
- Try to bring the pencils together:
  1. with one eye shut
  2. using both eyes
- About 5-10 % of humans are either stereo deficient or have no stereo perception. They still have depth perception though.

Stereopsis in biological vision

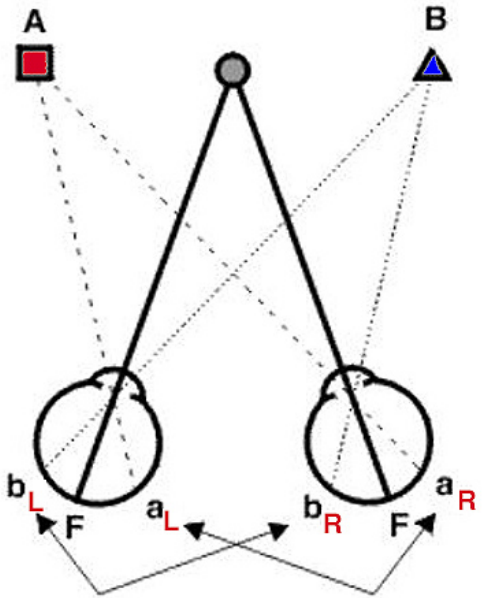
- Fusion: Disparate images from left and right eye are merged into a single unified percept.

  Only objects at approximately same distance as fixation target will be fused. Others give rise to “double vision” (diplopia).
- The word ”stereo” comes from the Greek word ”stereos” which means firm or solid.



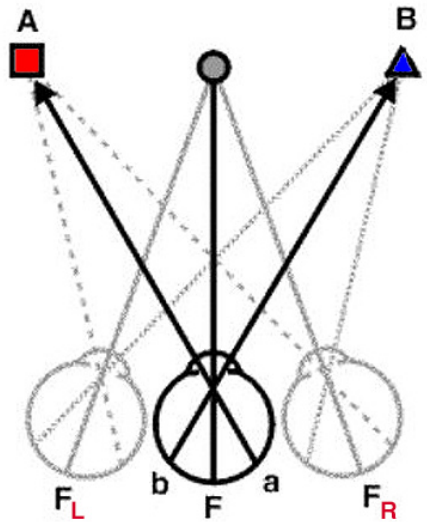
Stereo

- Corresponding retinal points (disparate points) are points stimulated on the retina that give rise to the same visual direction.
- Corresponding points have the same principle visual direction and non-corresponding points have different visual directions.



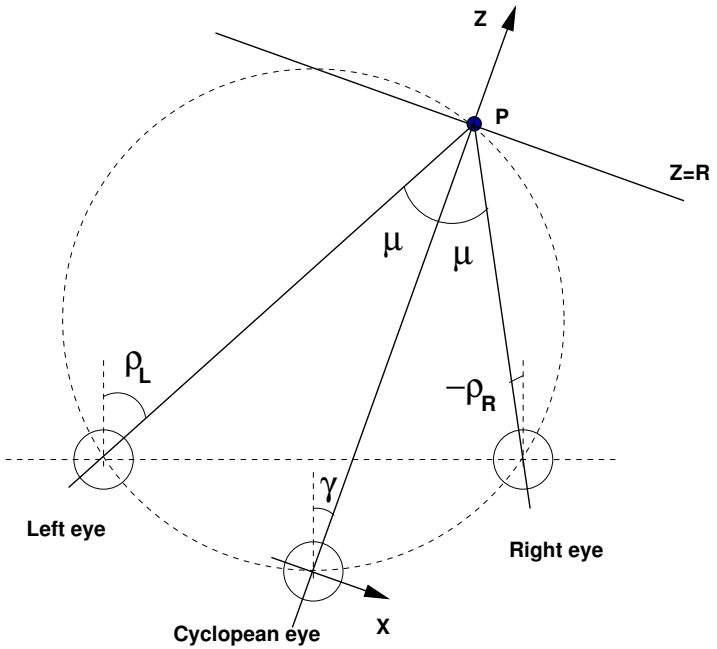
The cyclopean eye

- We see the world single and not double - binocular vision can be represented by a single eye, the cyclopean eye.
- The cyclopean eye is an imaginary eye situated midway between the two eyes.
- Used to determine directions of point A and B - point A stimulating temporal retina of right eye and nasal retina of left eye.



Stereo geometry: Verging cameras

- The principal rays of the two cameras converge at a point - the fixation point.
- In the plane defined by the fixation point and the centres of projection of the cameras, we have:



Stereo geometry

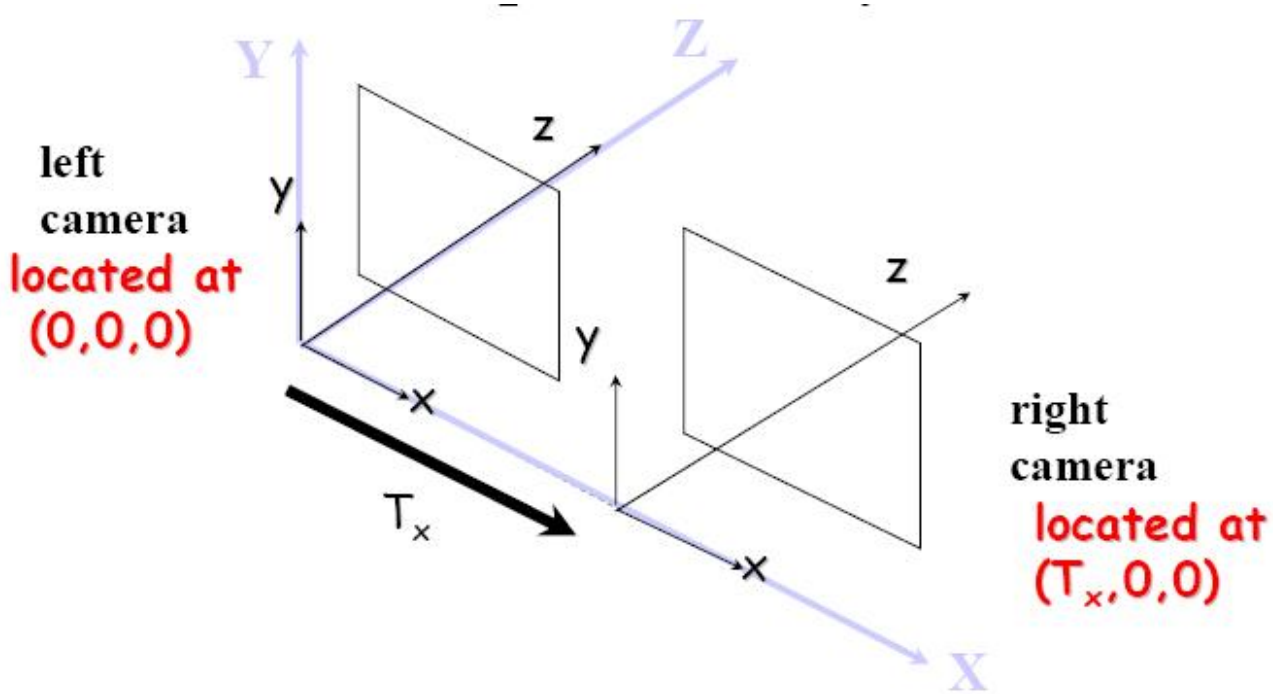
- Vergence angle 2μ - angle between principal directions.
- Gaze (version) angle γ - angle between primary direction and the ray from the cyclopean eye to the fixation point.
- Geometry:

$$\mu = \frac{1}{2}(\rho_L - \rho_R)$$

$$\gamma = \frac{1}{2}(\rho_L + \rho_R)$$

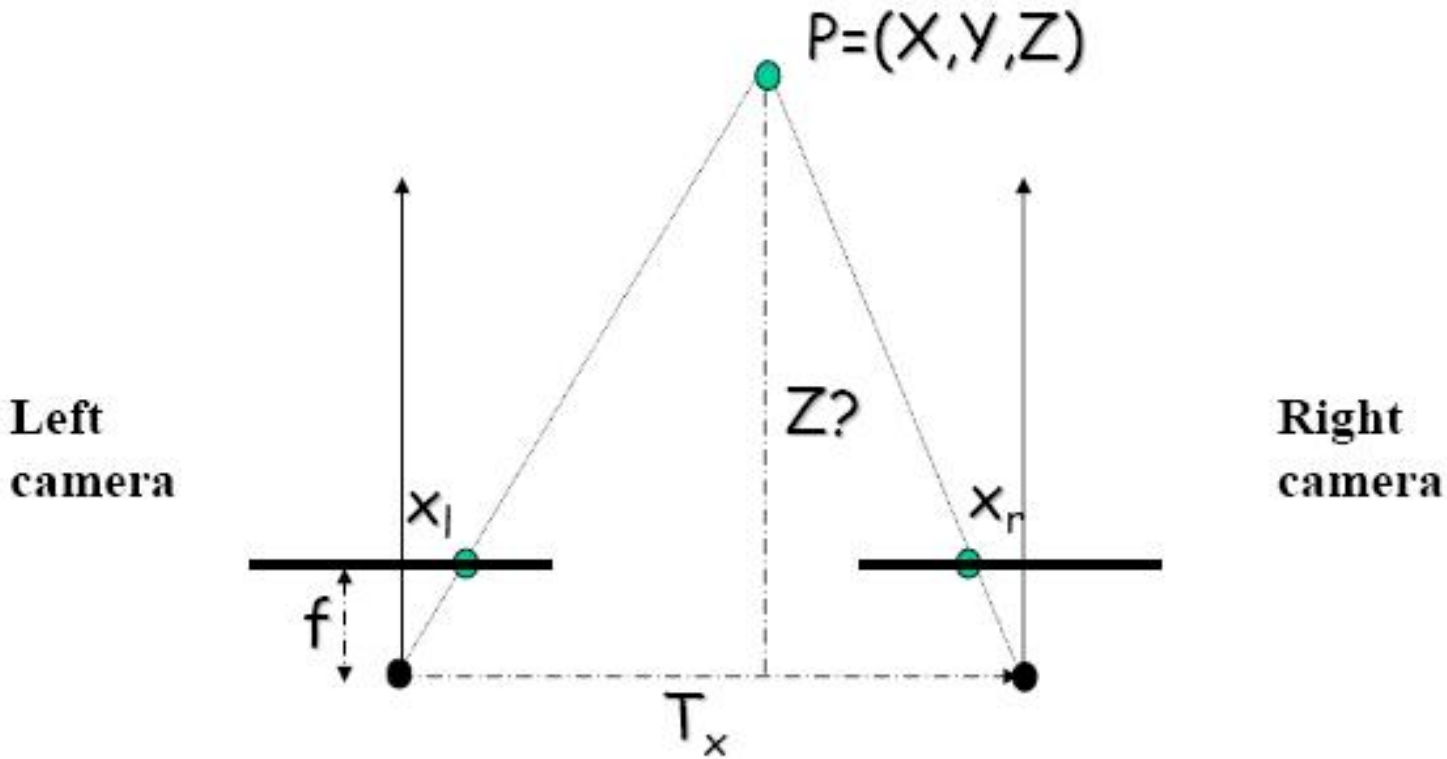
Parallel cameras

The optical axes are parallel to each other and perpendicular to the baseline connecting the optical axes.



Right camera is simply shifted by Tx units along the X axis. Otherwise, the cameras are identical (same orientation / focal lengths)

Top-down view



Translated by a distance T_x along X axis
(T_x is also called the stereo "baseline")

Parallel cameras

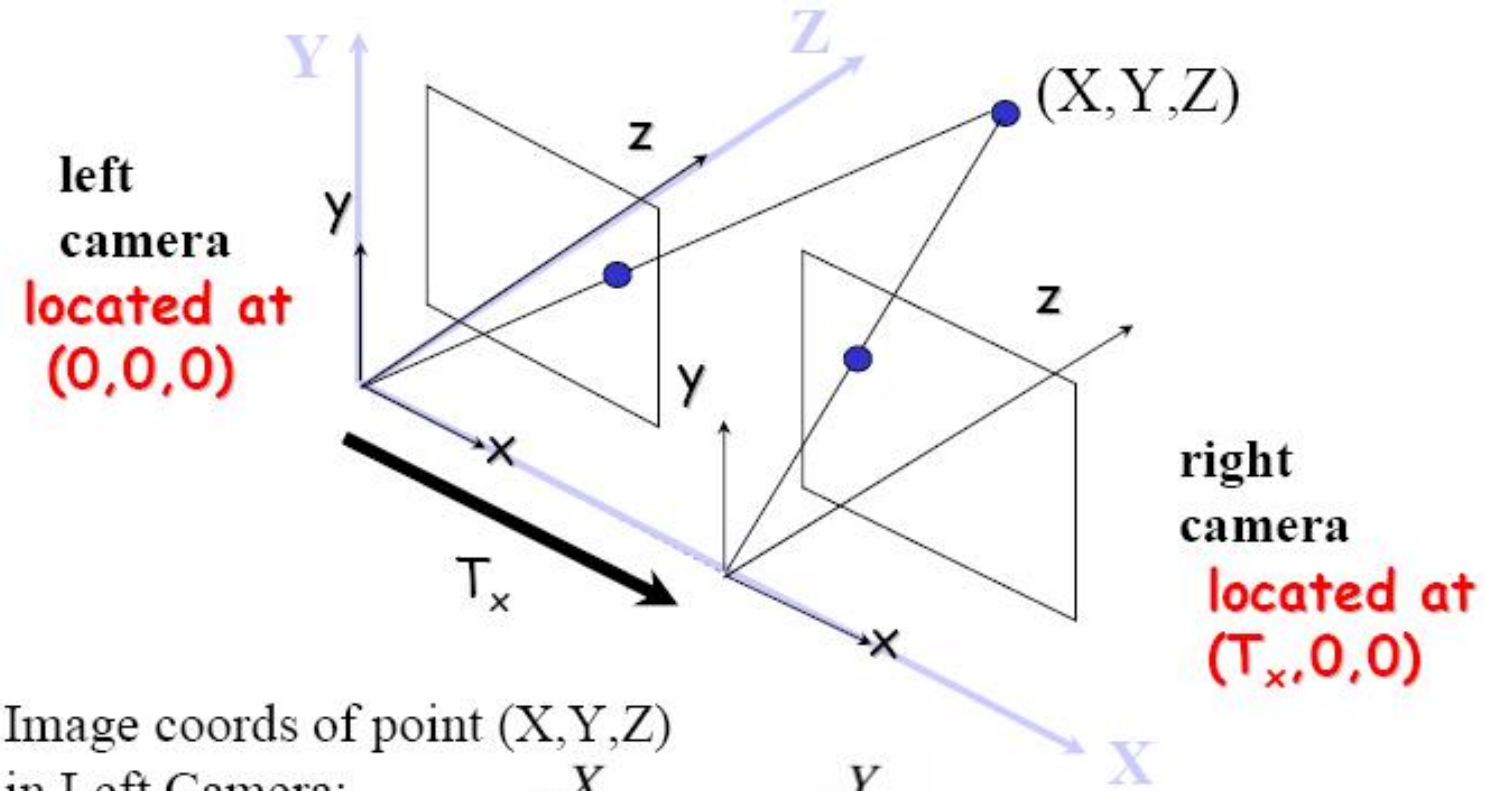
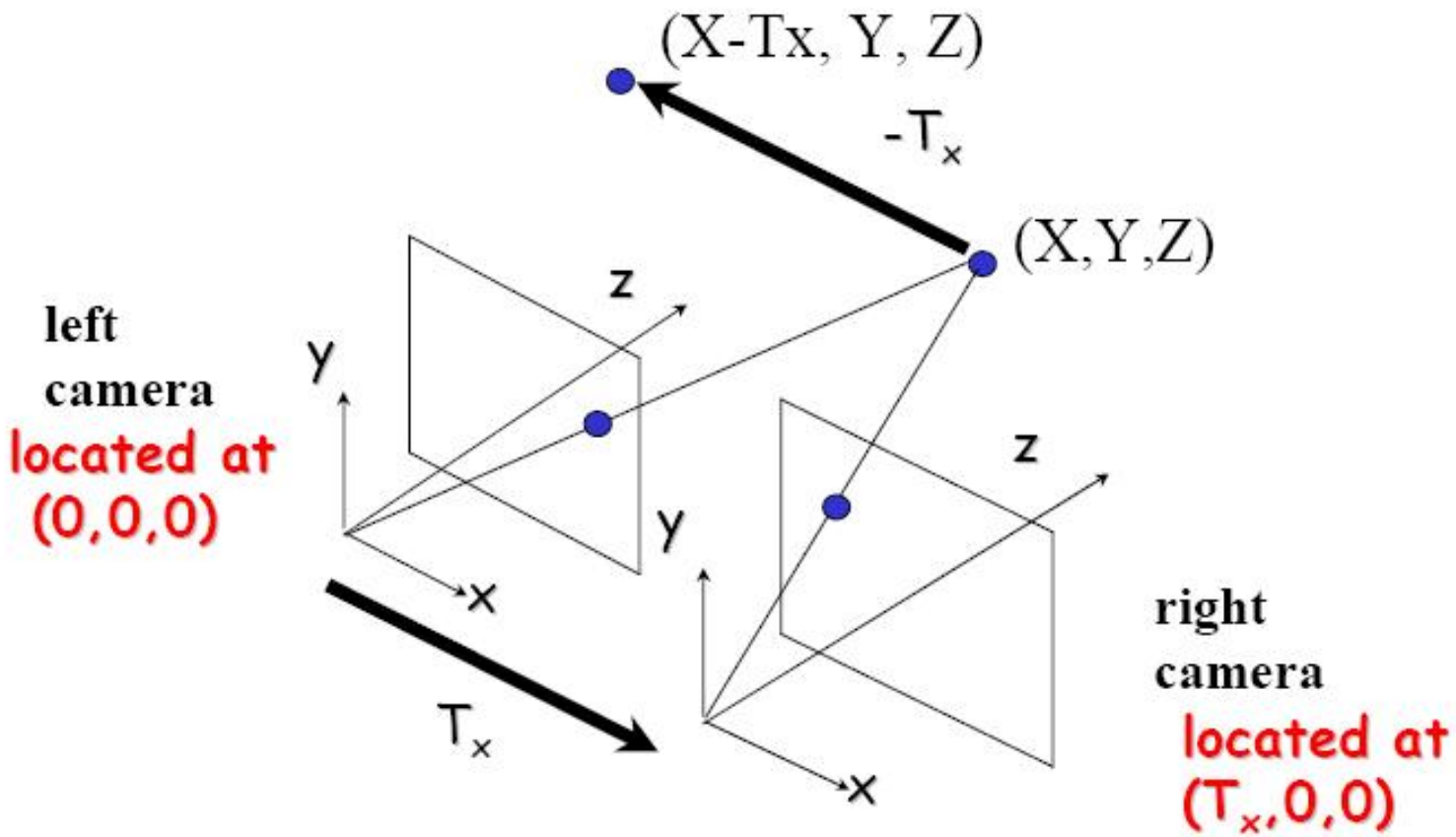


Image coords of point (X,Y,Z)
in Left Camera:

$$x_l = f\frac{X}{Z} \qquad y_l = f\frac{Y}{Z}$$

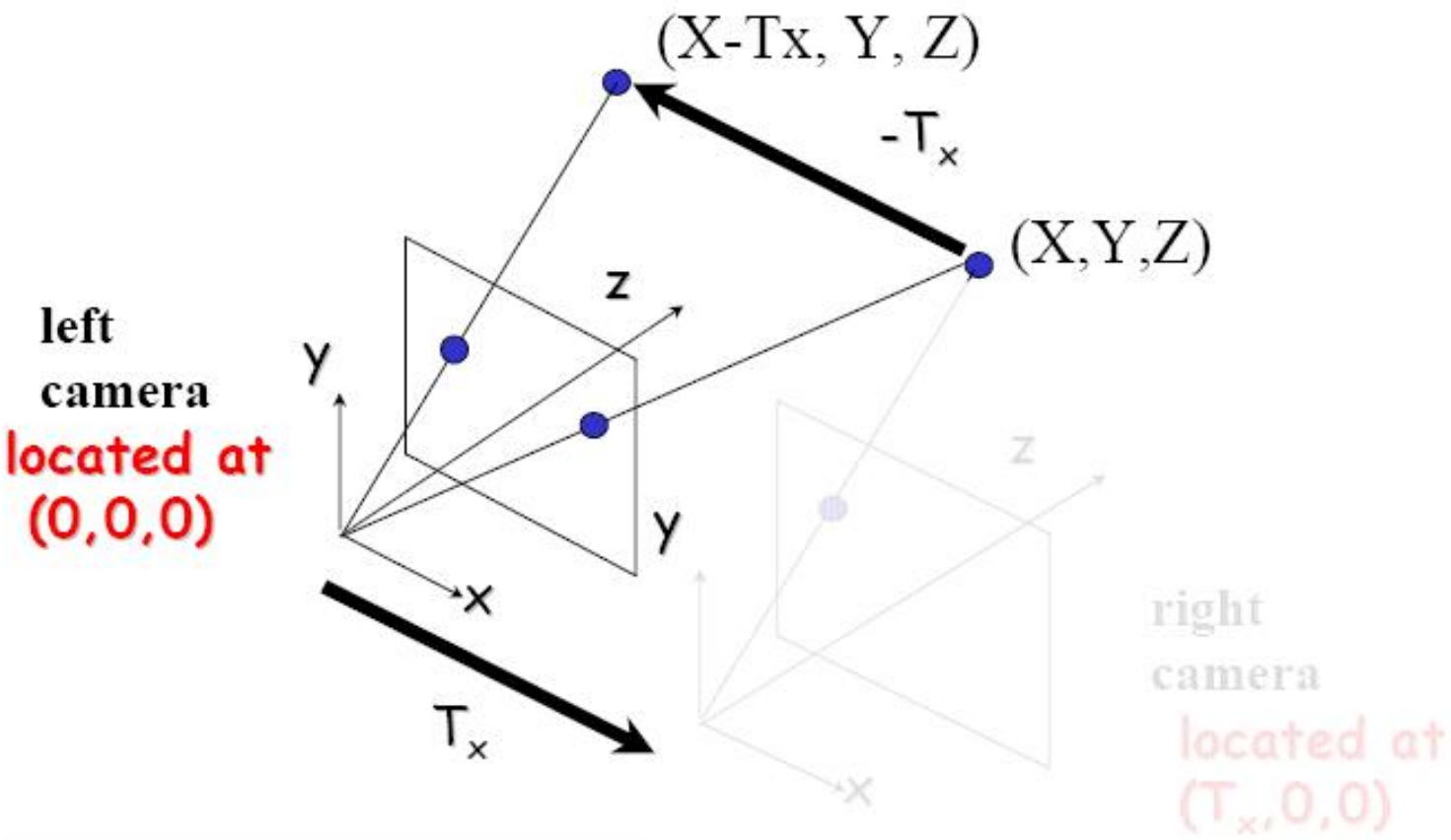
What are image coords of that same point
in the Right Camera?

Parallel cameras



Insight: translating camera to the right by Tx is equivalent to leaving the camera stationary and translating the world to the left by Tx.

Parallel cameras



$$x_r = f\frac{X - T_x}{Z} \qquad y_r = f\frac{Y}{Z}$$

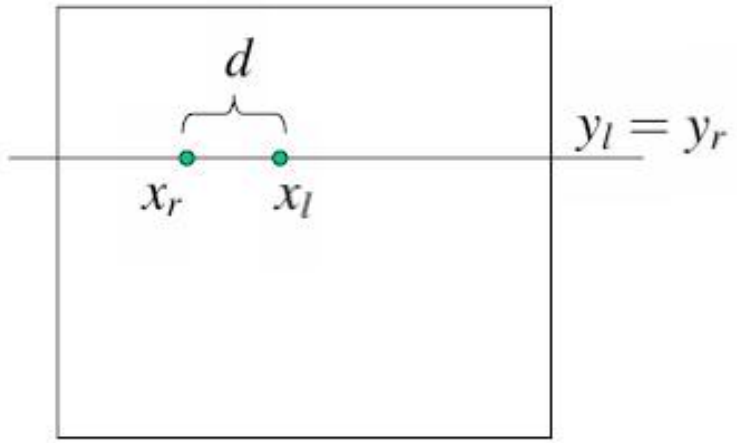
Parallel cameras

Left camera

$$x_l = f\frac{X}{Z} \qquad y_l = f\frac{Y}{Z}$$

Right camera

$$x_r = f\frac{X - T_x}{Z} \qquad y_r = f\frac{Y}{Z}$$



Stereo Disparity

$$d = x_l - x_r = f\frac{X}{Z} - \left(f\frac{X}{Z} - f\frac{T_x}{Z}\right)$$

$$d = \frac{f\,T_x}{Z}$$

depth **baseline**

$$Z = \frac{f\,T_x}{d}$$

disparity

Important equation!

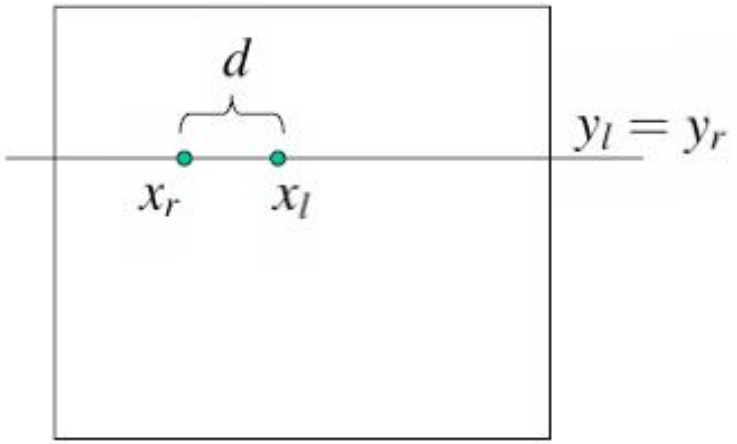
Parallel cameras

Left camera

$$x_l = f\frac{X}{Z} \qquad y_l = f\frac{Y}{Z}$$

Right camera

$$x_r = f\frac{X - T_x}{Z} \qquad y_r = f\frac{Y}{Z}$$



Note: Depth and stereo disparity are inversely proportional

depth

$$Z = \frac{f\,T_x}{d}$$

disparity

Important equation!

Parallel cameras

- In a cyclopean coordinate system with the x-axis along the baseline T and the z-axis parallel to the optical axes, we have

$$\frac{x_L}{f_L} = \frac{X + T/2}{Z}, \qquad \frac{x_R}{f_R} = \frac{X - T/2}{Z}$$

$$\frac{y_L}{f_L} = \frac{Y}{Z}, \qquad \frac{y_R}{f_R} = \frac{Y}{Z}$$

Reconstruction from disparities

$$X + \frac{T}{2} - \frac{x_L}{f} Z = 0 \tag{1}$$

$$X - \frac{T}{2} - \frac{x_R}{f} Z = 0 \tag{2}$$

$$(2) - (1) \Rightarrow \frac{(x_L - x_R)}{f} Z = T \Rightarrow Z = \frac{Tf}{x_L - x_R}$$

$$x_L \cdot (2) - x_R \cdot (1) \Rightarrow (x_L - x_R) X = \frac{T}{2}(x_L + x_R) \Rightarrow X = \frac{T}{2} \frac{(x_L + x_R)}{(x_L - x_R)}$$

Depth Z is inversely proportional to disparity $d = (x_L - x_R)$ and proportional to the baseline t.

Disparity

- Let (x_L, y_L) and (x_R, y_R) be corresponding image points.
- The difference between these coordinates is called disparity.
  - Horizontal disparity: $x_L - x_R$
  - Vertical disparity: $y_L - y_R$
- Note: For parallel cameras the vertical disparity is zero and it is sufficient to perform matching along a horizontal line.

Trade-off for wider baselines

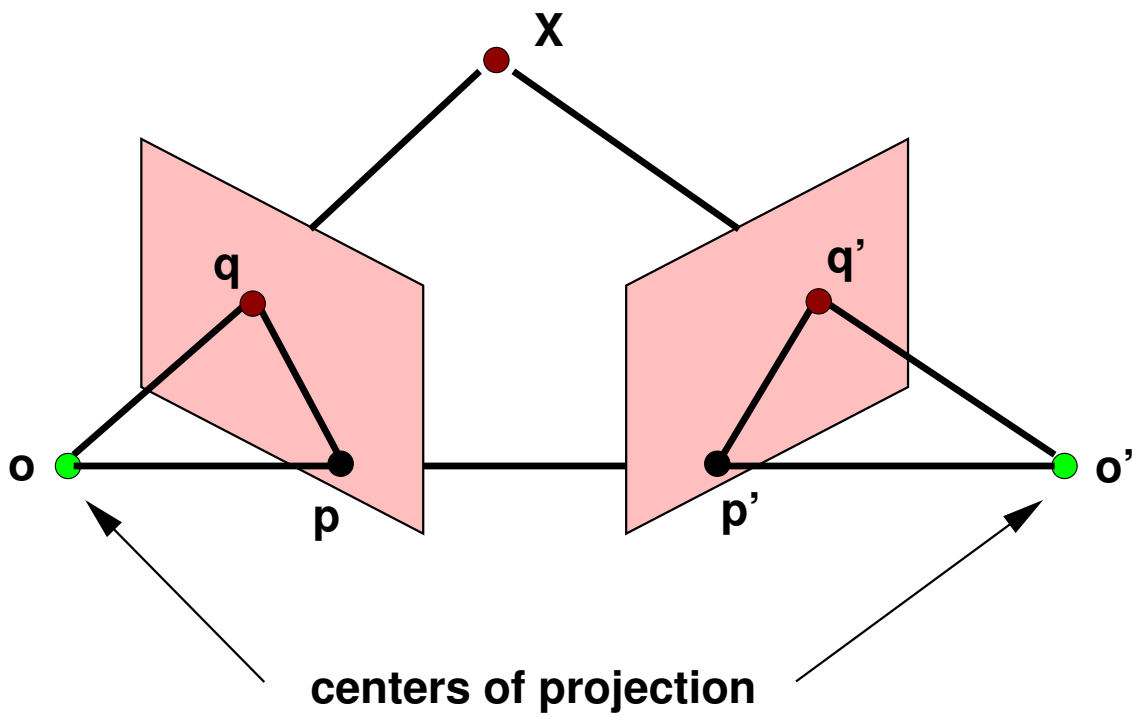
- Differentiate

$$Z = \frac{Tf}{d}$$

$$\Rightarrow \frac{\delta Z}{\delta d} = -\frac{Tf}{d^2} = -\frac{Z^2}{Tf}$$

- The depth errors depends quadratically on the depth.
- Large baseline stereo gives
  - more accurate reconstruction, but
  - harder matching problems (larger perspective distortions + larger proportion of occluded regions).

General case: Epipolar geometry

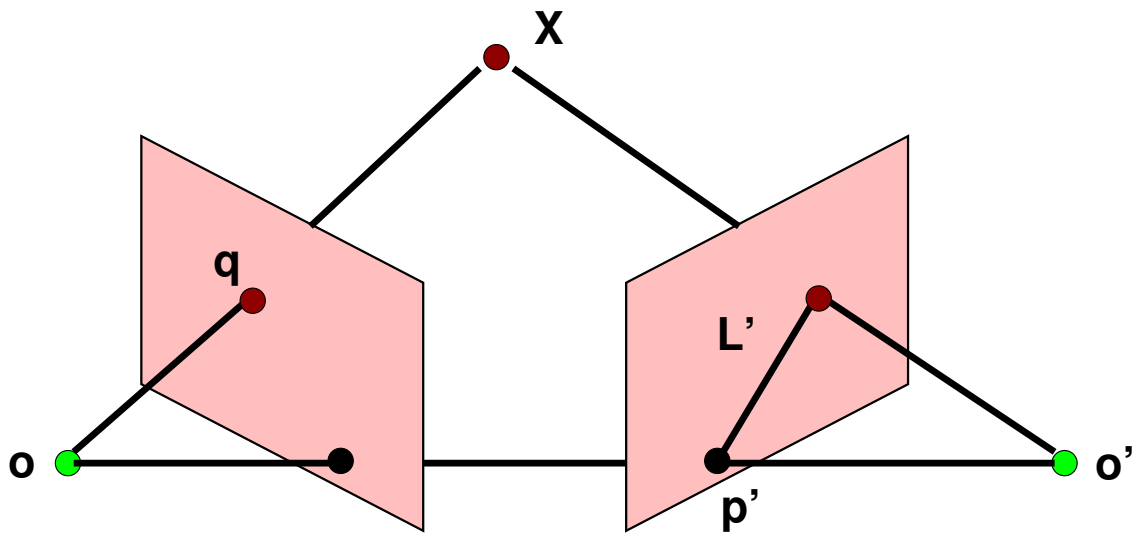


Epipolar plane: A plane through a point X and the optical centers.

The projection of the optical center $o'(o)$ is called the epipole $p(p')$.

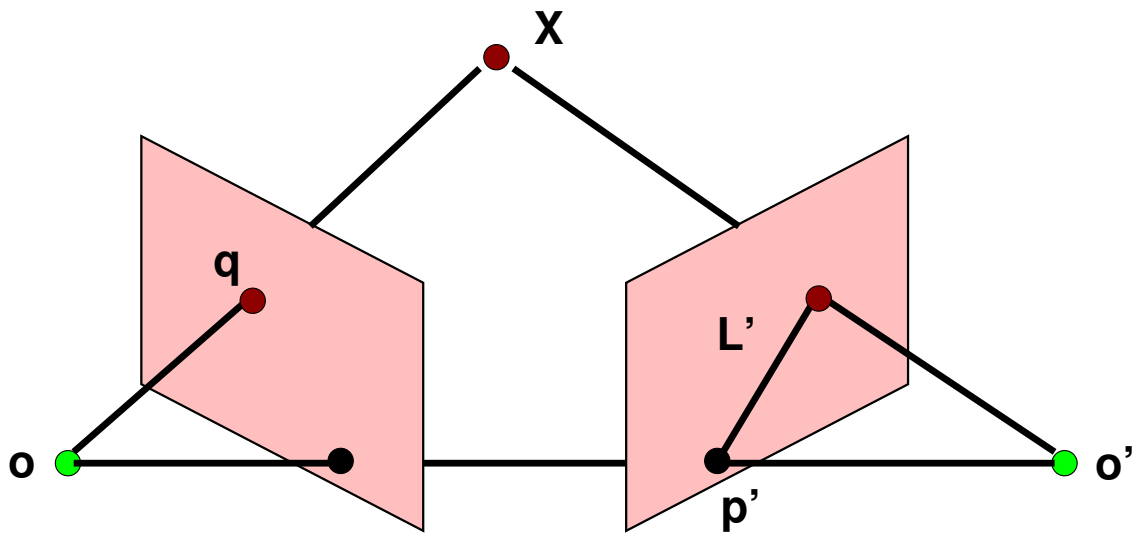
Note: Any epipolar plane projects to a line that goes through the epipole. Such a line is called an epipolar line.

The epipolar transform



- Given a point q in one image, the corresponding image point q' must be in the epipolar plane through o, o' and q.
- This plane projects to an epipolar line through p' and q' is constrained to be on this line.
- Thus: It is enough to perform a one-dimensional search for the matching point.

The epipolar transform



- The map from points in one image plane to epipolar lines in the other image plane is called the epipolar transform:

$$q \rightarrow L'(q)$$

The essential matrix

- Assume the two cameras are related by:

  R - a relative rotation, and

  t - a relative translation (baseline).
- The normal of the epipolar plane given by a point q is

$$n = t \times q$$

- Expressed in the right camera's coordinate system the normal is

$$n' = R\, n = R\, (t \times q)$$

- The point q' lies on the epipolar plane if

$$q'^T n' = 0 \rightarrow q'^T R\, (t \times q) = 0$$

The essential matrix

- This so called epipolar constraint equation can be written as

$$q'^T E q = 0$$

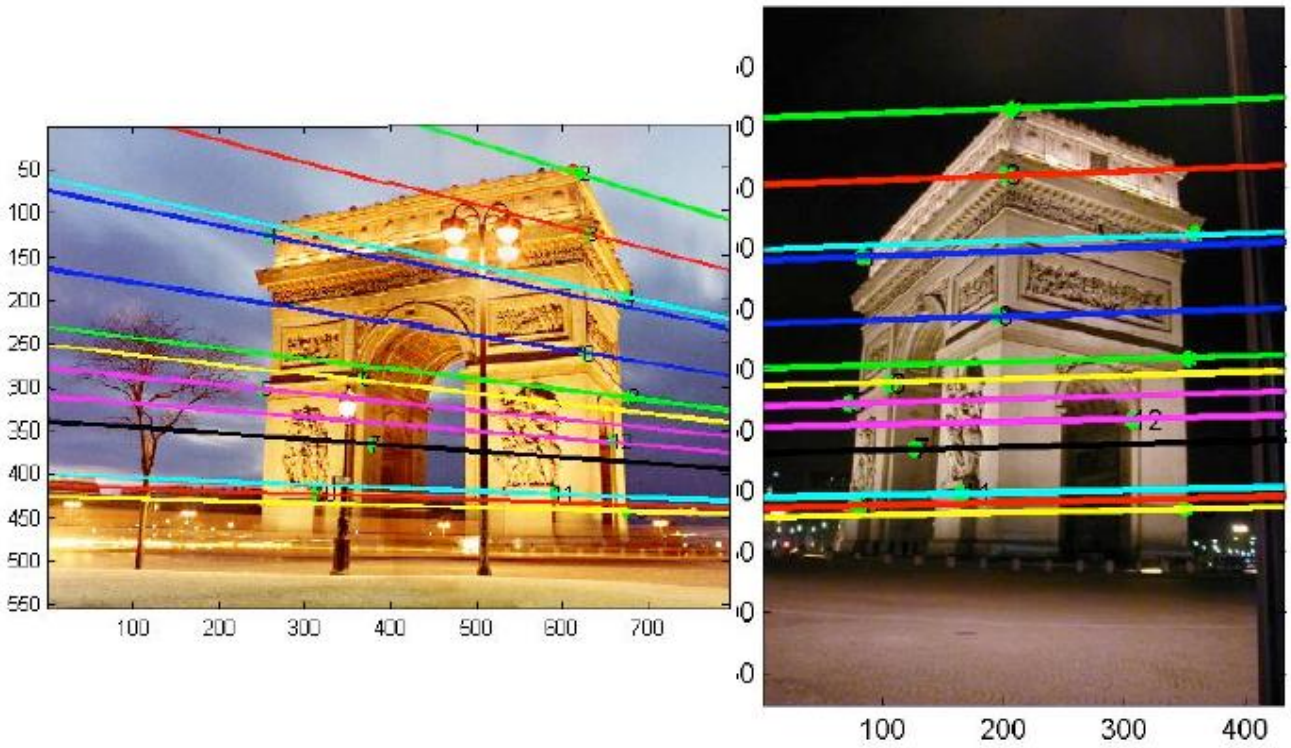
where the matrix $E = RT_t$ is called the essential matrix, and

$$T_t = \begin{pmatrix} 0 & t_z & -t_y \\ -t_z & 0 & t_x \\ t_y & -t_x & 0 \end{pmatrix}$$

is a skew-symmetric matrix that has the same effect as a cross product with the baseline t.

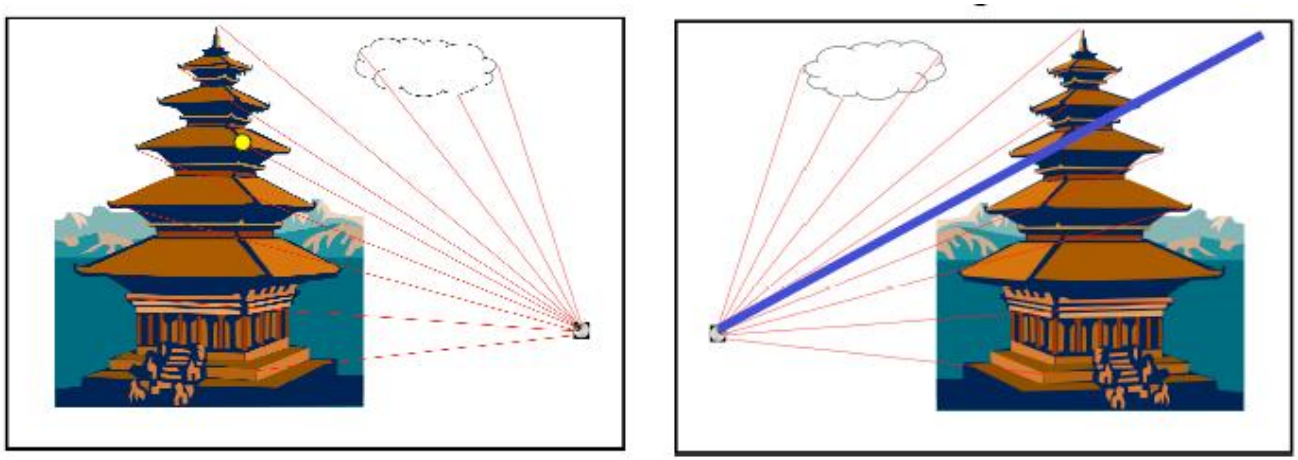
The essential matrix

- The essential matrix $E = RT_t$ encodes epipolar line constraints.
- In the image the normal $n' = Eq$ can be seen as the epipolar line corresponding to q.
- In the other image the line equation is: $q'^T n' = 0$.



Essential and Fundamental matrices

- The essential and fundamental matrices are 3×3 matrices that encode the epipolar geometry of two views.
  - Essential matrix E relates image coordinates
  - Fundamental matrix $F = K'^{-T} E K^{-1}$ relates pixel coordinates
  - K and K' are 3×3 camera matrices with intrinsic parameters.
- Usage: Given a point in one image, multiplying by the essential or fundamental matrix will tell us which epipolar line to search along in the second view.



Determining the essential matrix

- Use features (e.g. SIFT) matched between the two images.
- The essential matrix depends on
  - Rotation R (3 parameters)
  - Translation t (3 parameters)
- The matrix is, however, homogeneous in components of t

  $\Rightarrow$ Totally 5 unknowns.
- Each correspondence gives one constraint $\Rightarrow$ 5 matches needed.

Bundle adjustment: multiple cameras

- Assume you have K camera images with unknown projection matrices M_k and projections q_{ik} of N unknown 3D points Q_i.
- Then you can set up a large system of equations and search for M_k and Q_i by minimizing

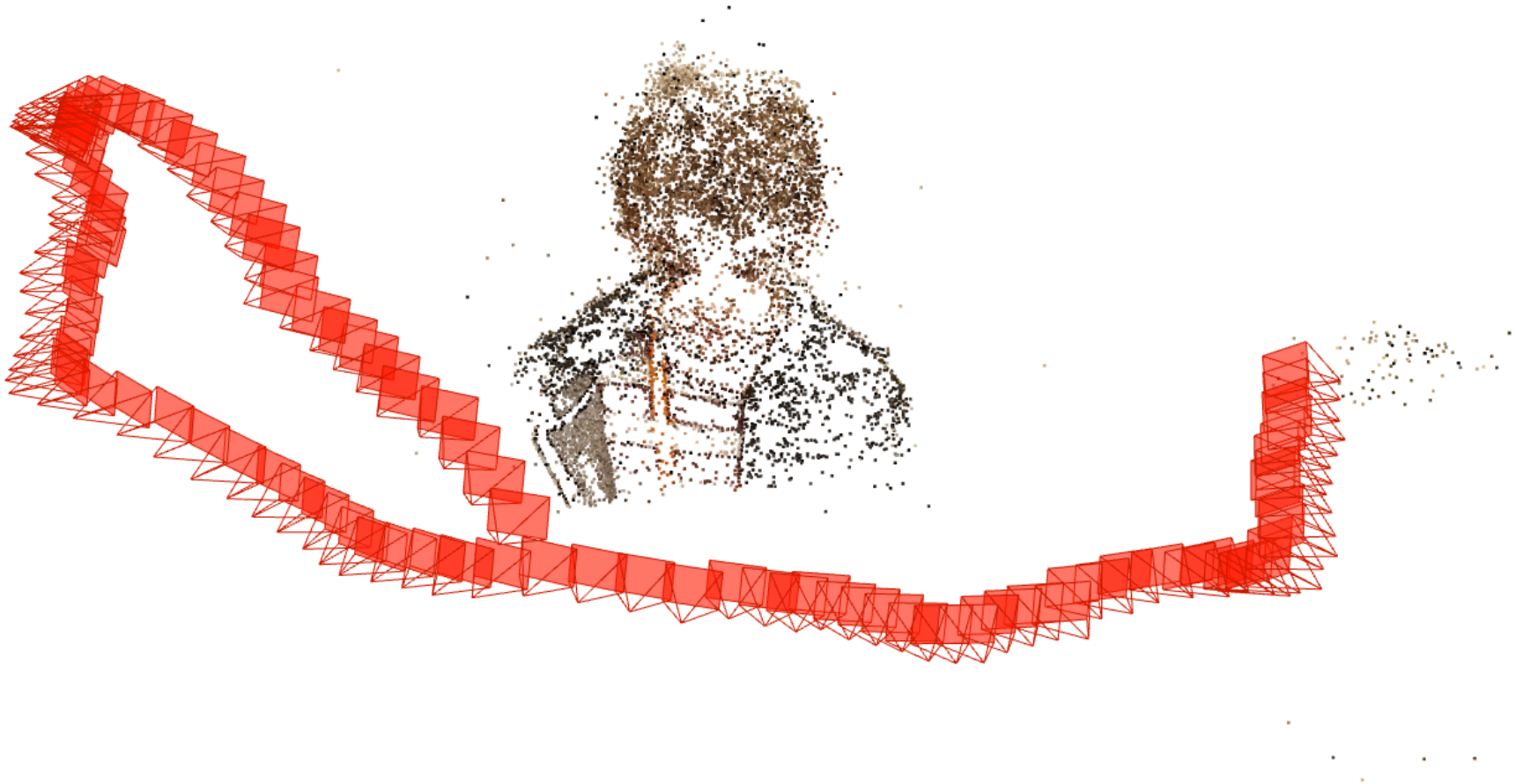
$$\min_{\{M_k, Q_i\}} \sum_{k=1}^{K} \sum_{i=1}^{N} d(q_{ik}, f(M_k, Q_i))$$

where

$$d(q_{ik}, f(M_k, Q_i)) = \left(x_{ik} - \frac{M_k^1 Q_i}{M_k^3 Q_i} \right)^2 + \left(y_{ik} - \frac{M_k^2 Q_i}{M_k^3 Q_i} \right)^2$$

- After initialization by first computing the essential matrices between pairs of camera images, this can be solved iteratively.

Bundle adjustment: example

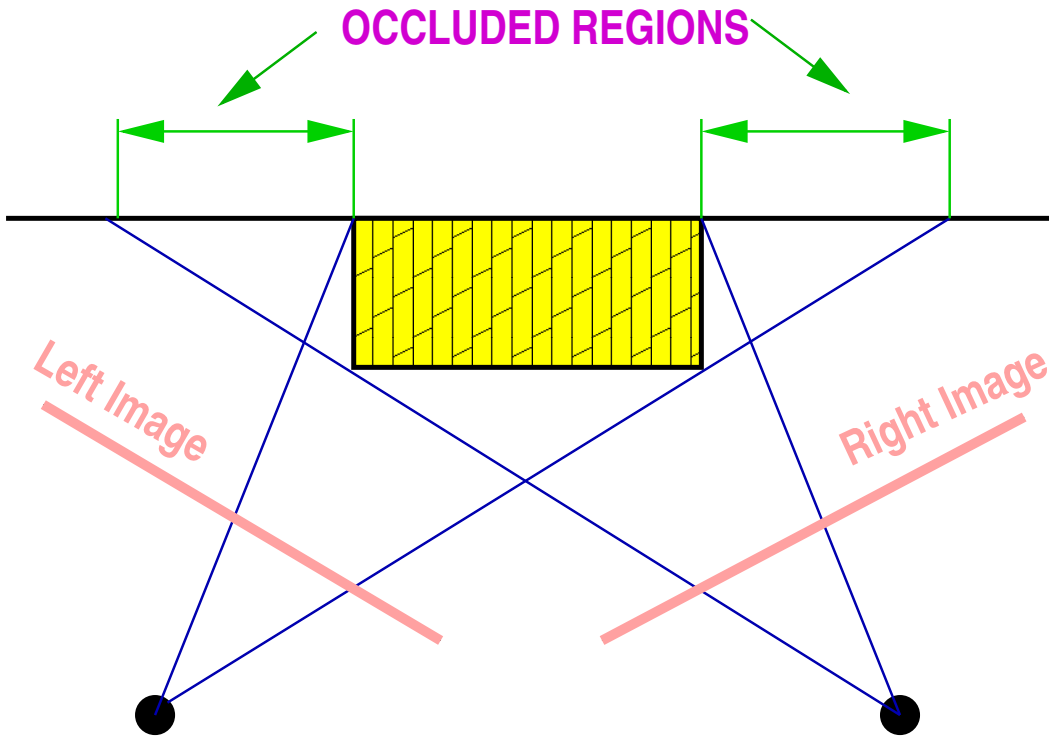


Feature extraction, matching and bundle adjustment using `ColMap`.

Establishing correspondence - Problems

Non-trivial problem, because of several reasons.

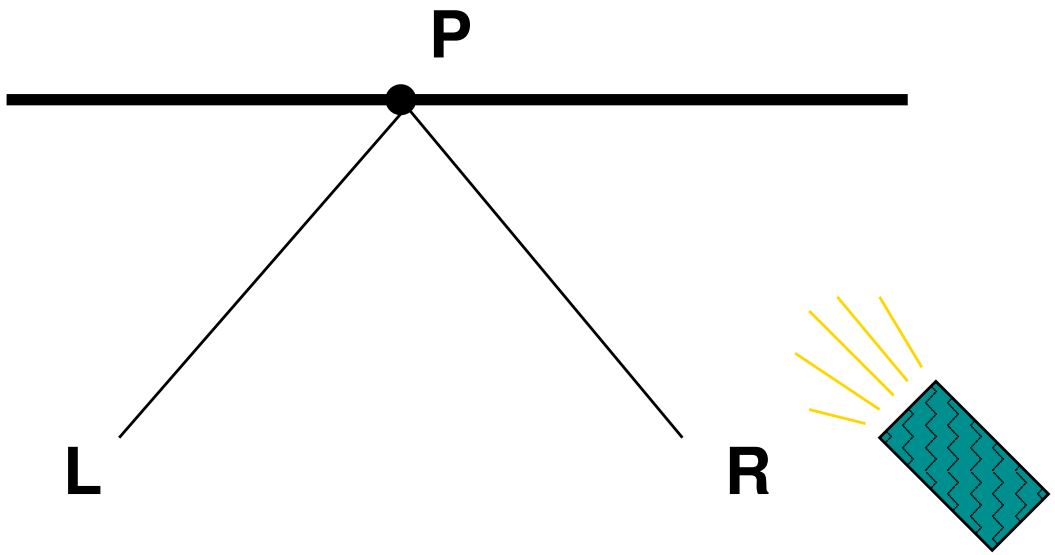
- Occlusions: some parts of scene are only seen by one camera.



Establishing correspondence - Problems

- Brightness variations: in general, cameras often have different orientations relative to the source of illumination.

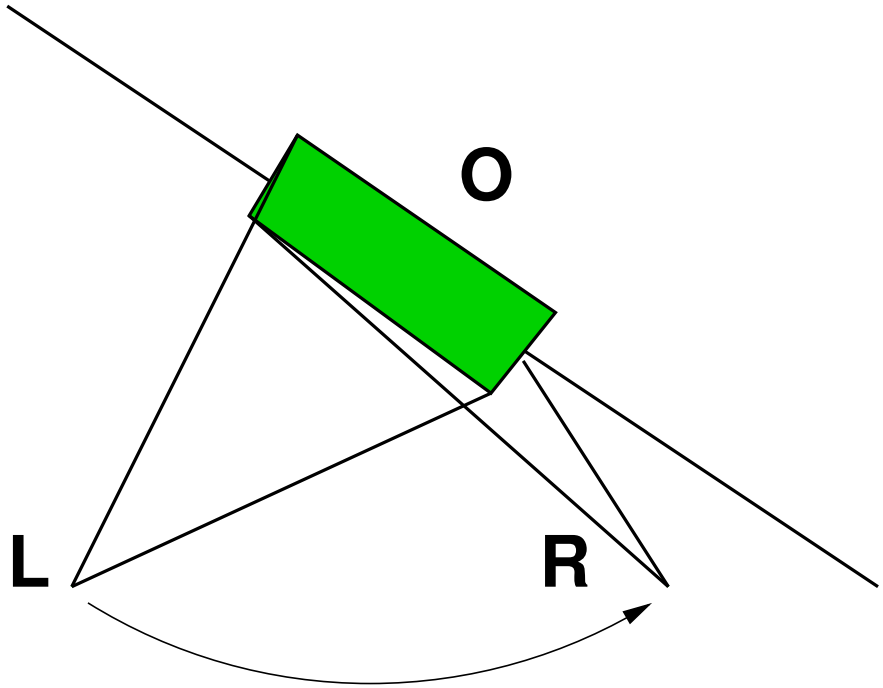
  $\Rightarrow$ Brightness will not be the same in the two regions.



P appears brighter to L than to R.

Establishing correspondence - Problems

- Distortions due to perspective effects.

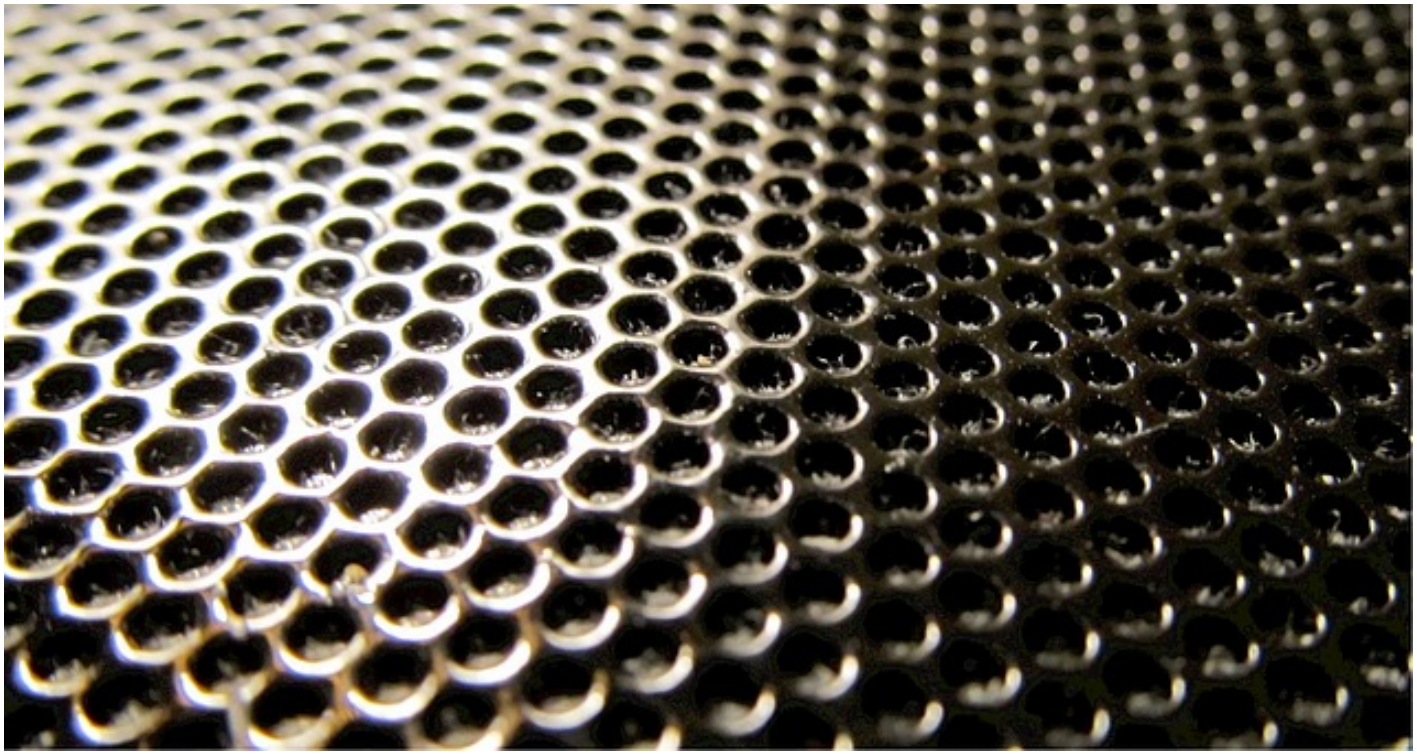


foreshortening

O appears to be larger to L than R.

Establishing correspondence - Problems

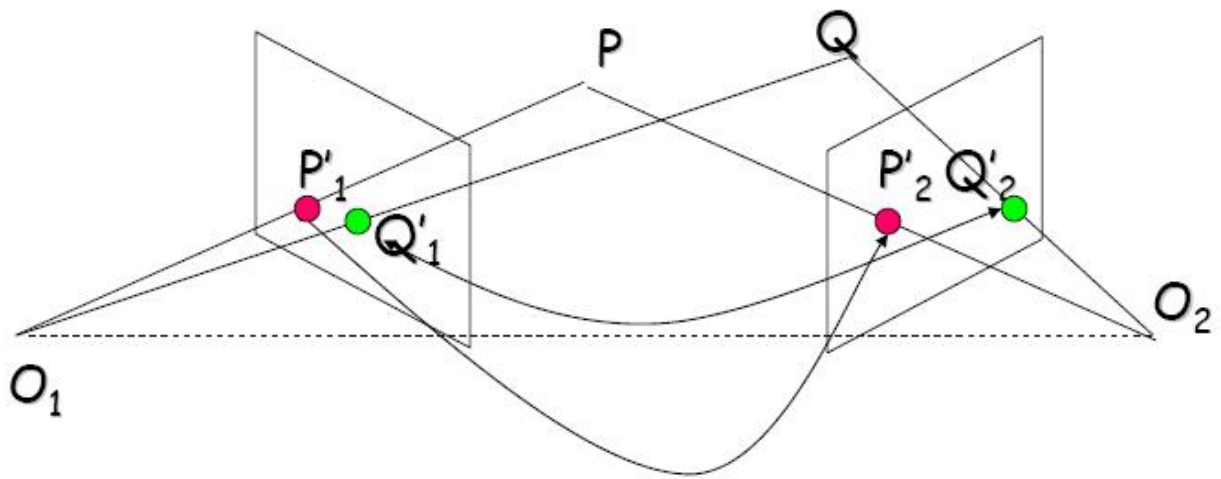
- Repetitive textures $\Rightarrow$ ambiguous matches.



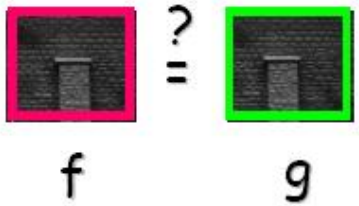
Lack of unique surface markings.

Monotonic ordering constraint

- If we assume that all 3D points lie on the same surface and are visible by both eyes, then points on the same epipolar line are in the same order in both eyes.
- This assumption does not always hold (test by holding two fingers in-front of each other).



Comparing Windows:

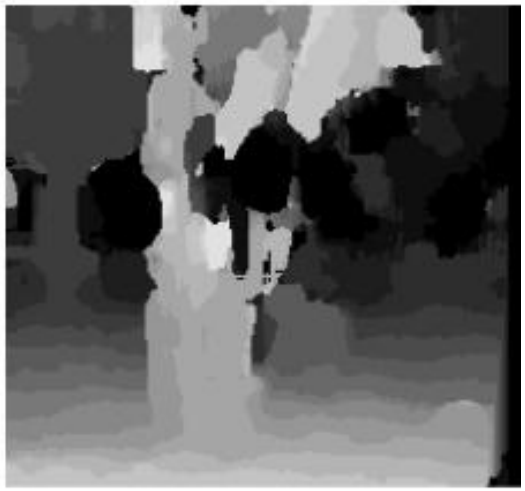


$$SSD = \sum_{[i,j]} (f(i,j) - g(i,j))^2$$ most natural

$$SAD = \sum_{[i,j]} ||f(i,j) - g(i,j)||$$ often more robust

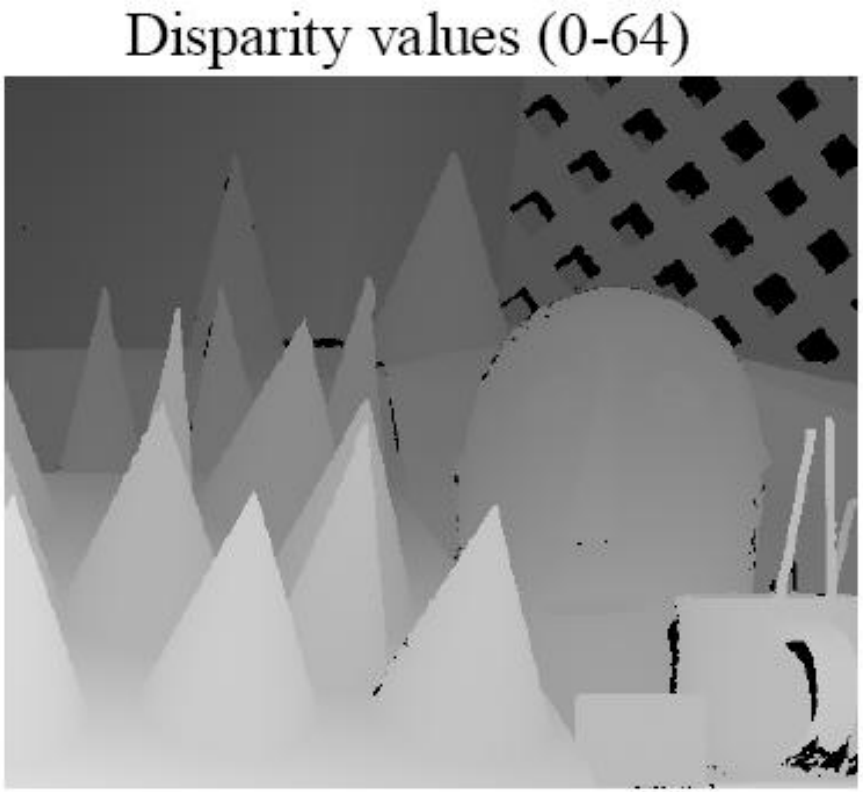
Results using different window sizes

W = 3



W = 20

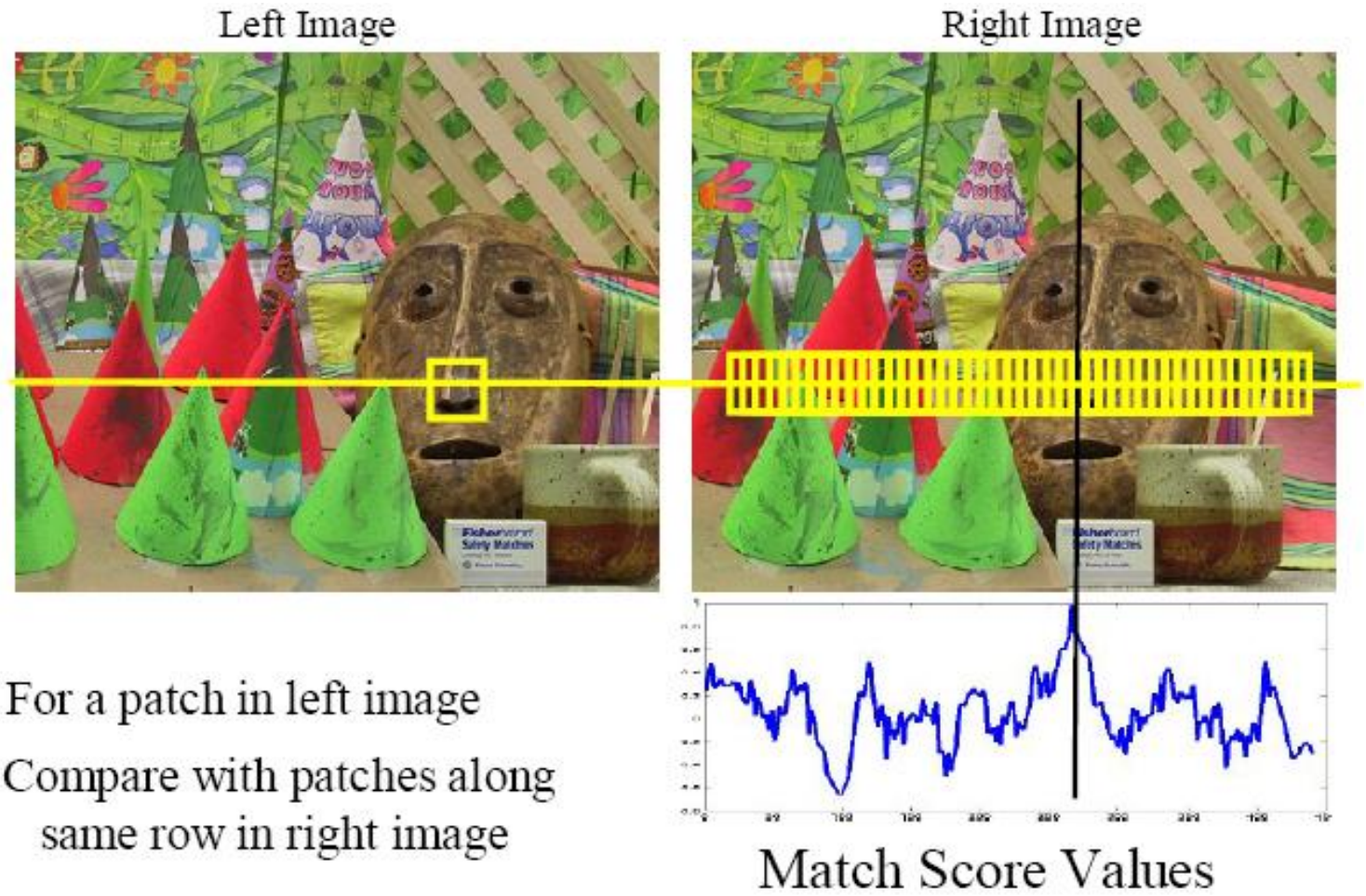
Example: Disparity map



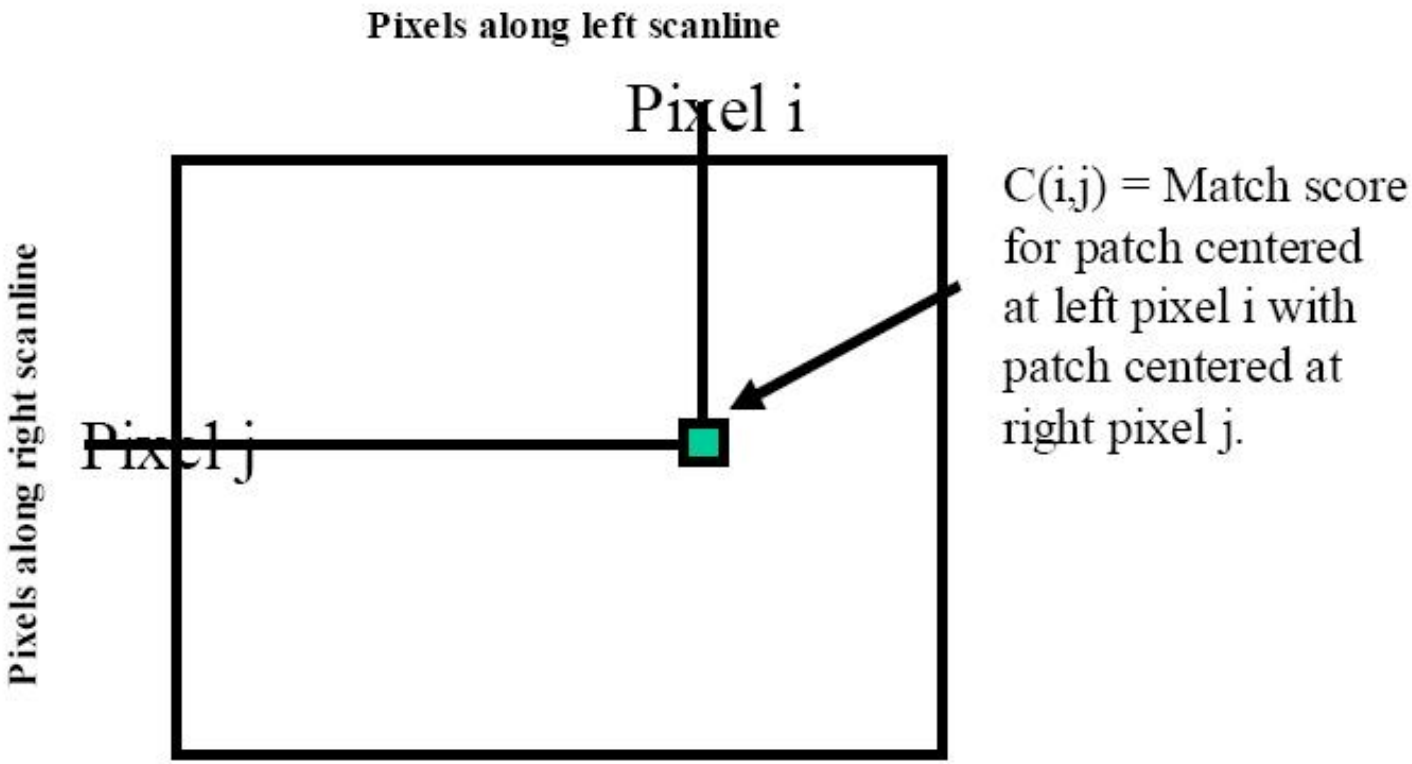
Note how disparity is larger (brighter) for closer surfaces.

Common benchmark example – far too easy compared to real world.

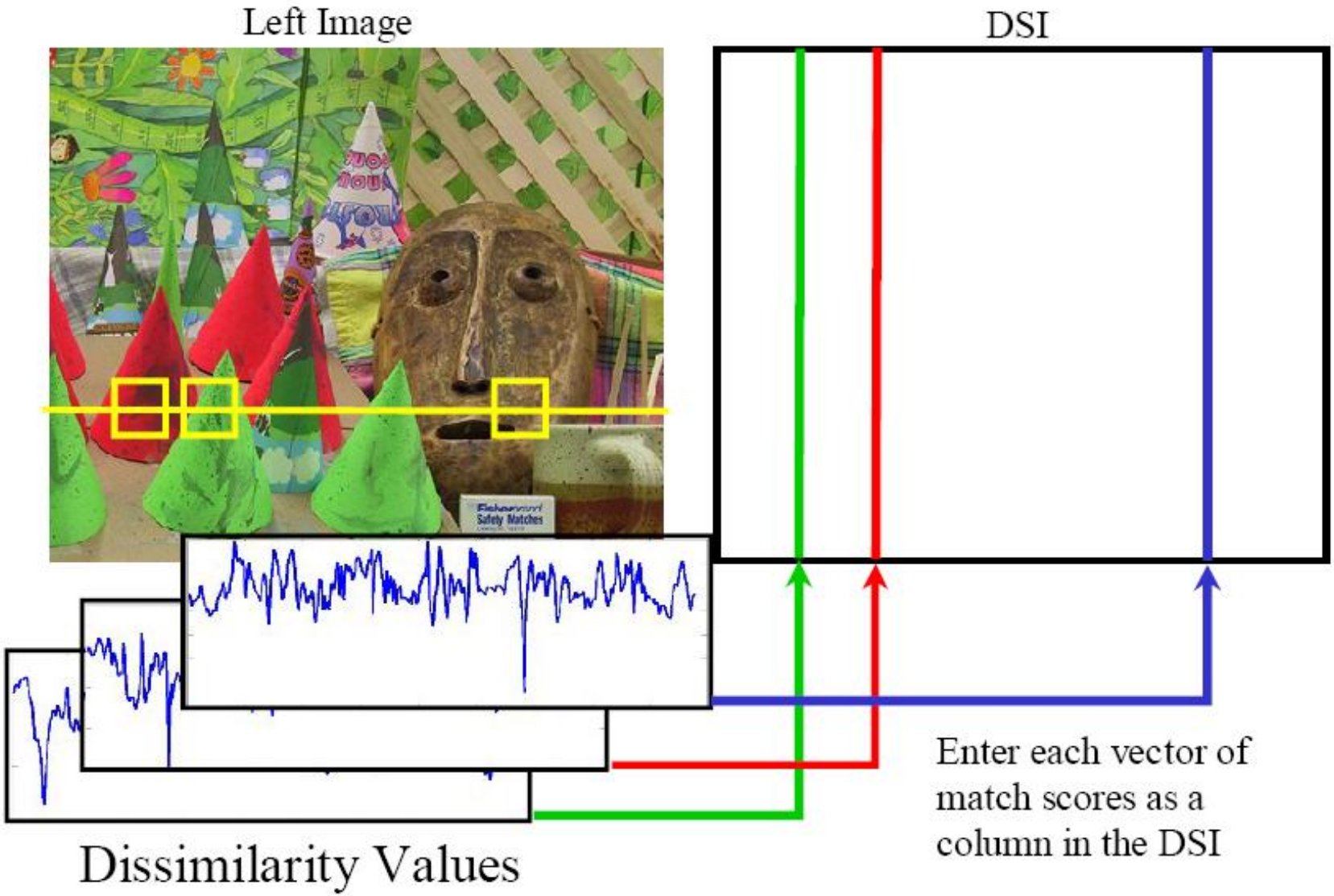
Example: Disparity map



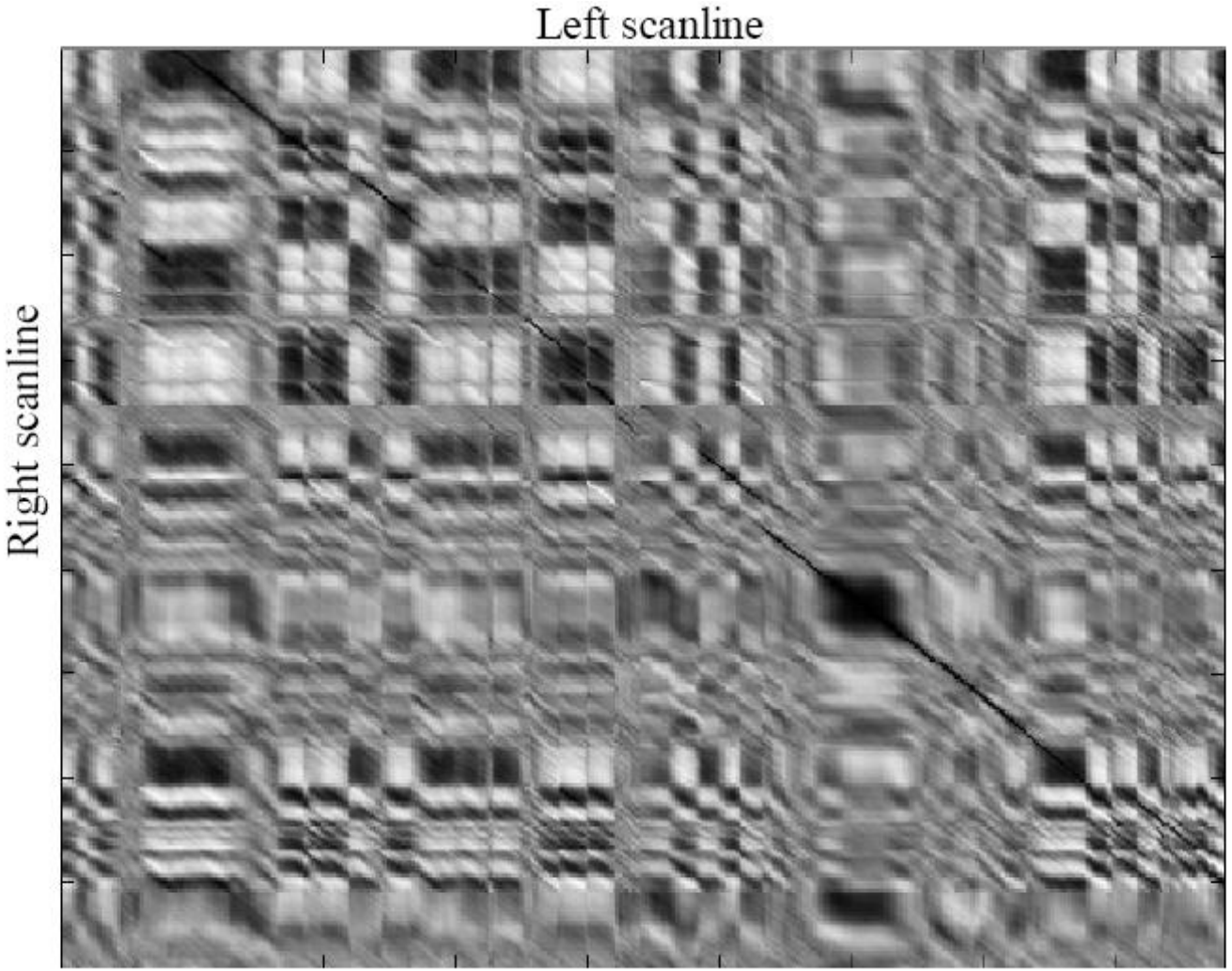
Disparity Space Image



Disparity Space Image

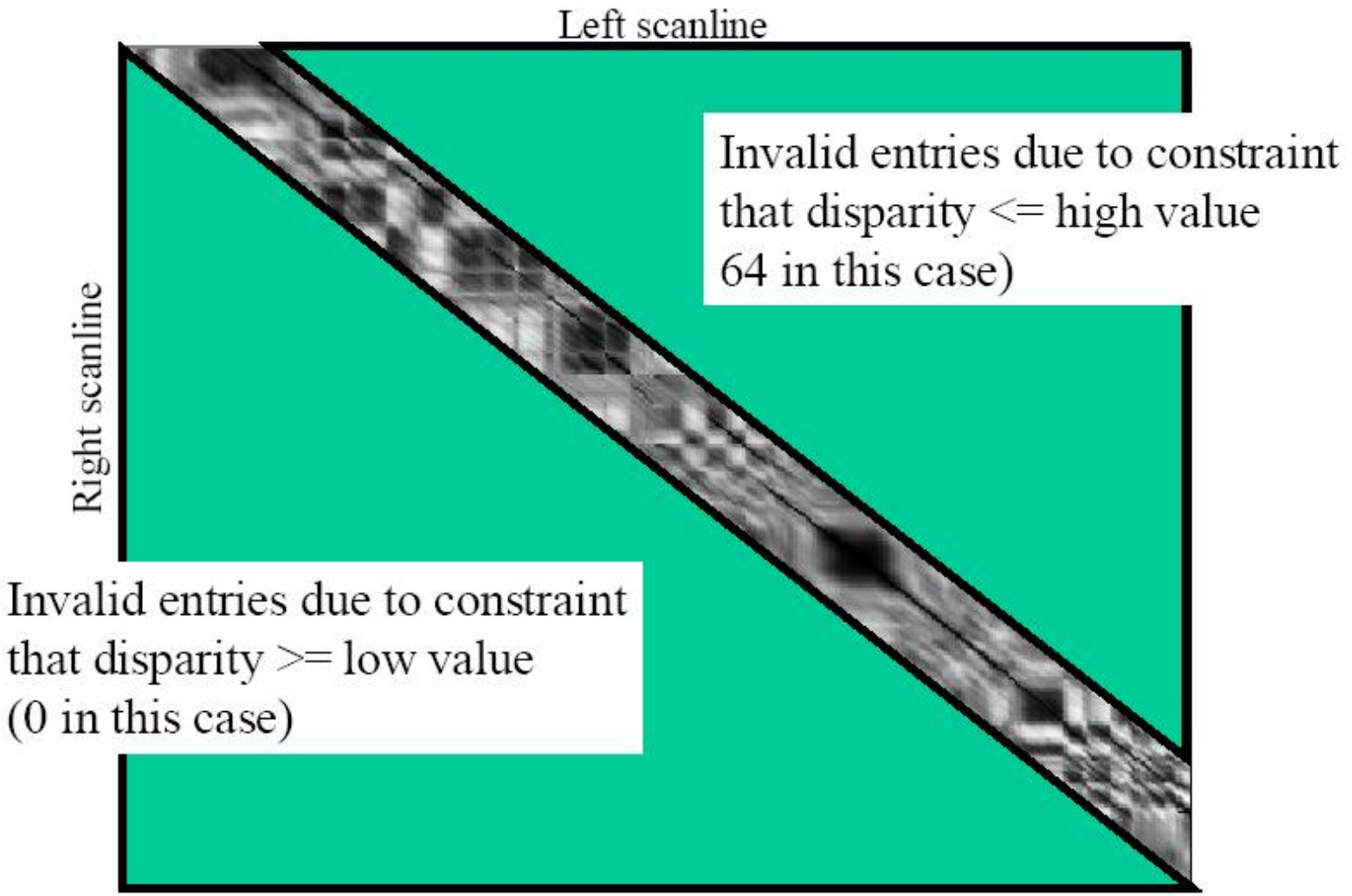


Disparity Space Image



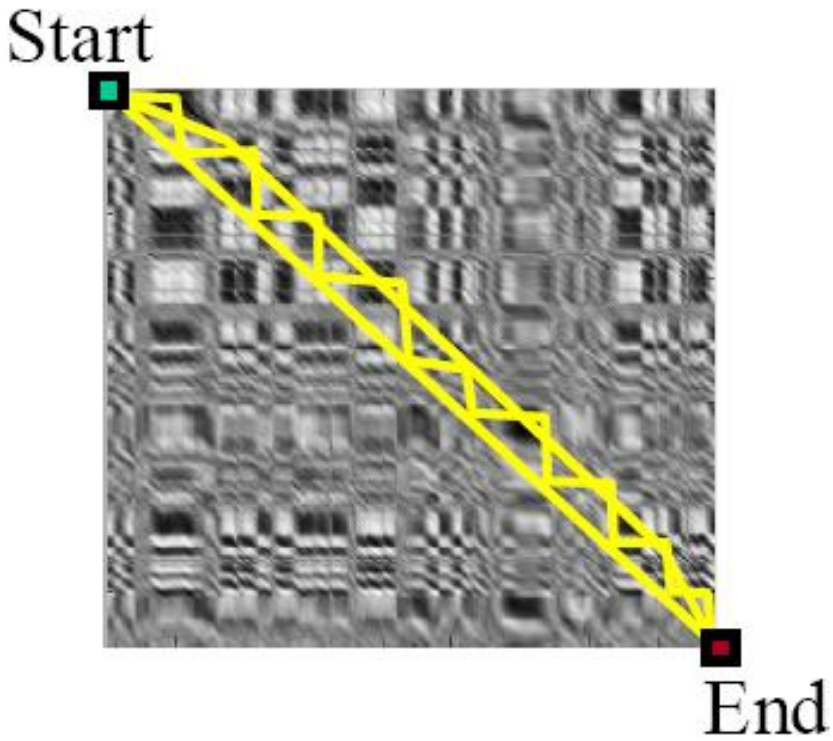
Observe the diagonal black lines corresponding to true matches.

Disparity Space Image



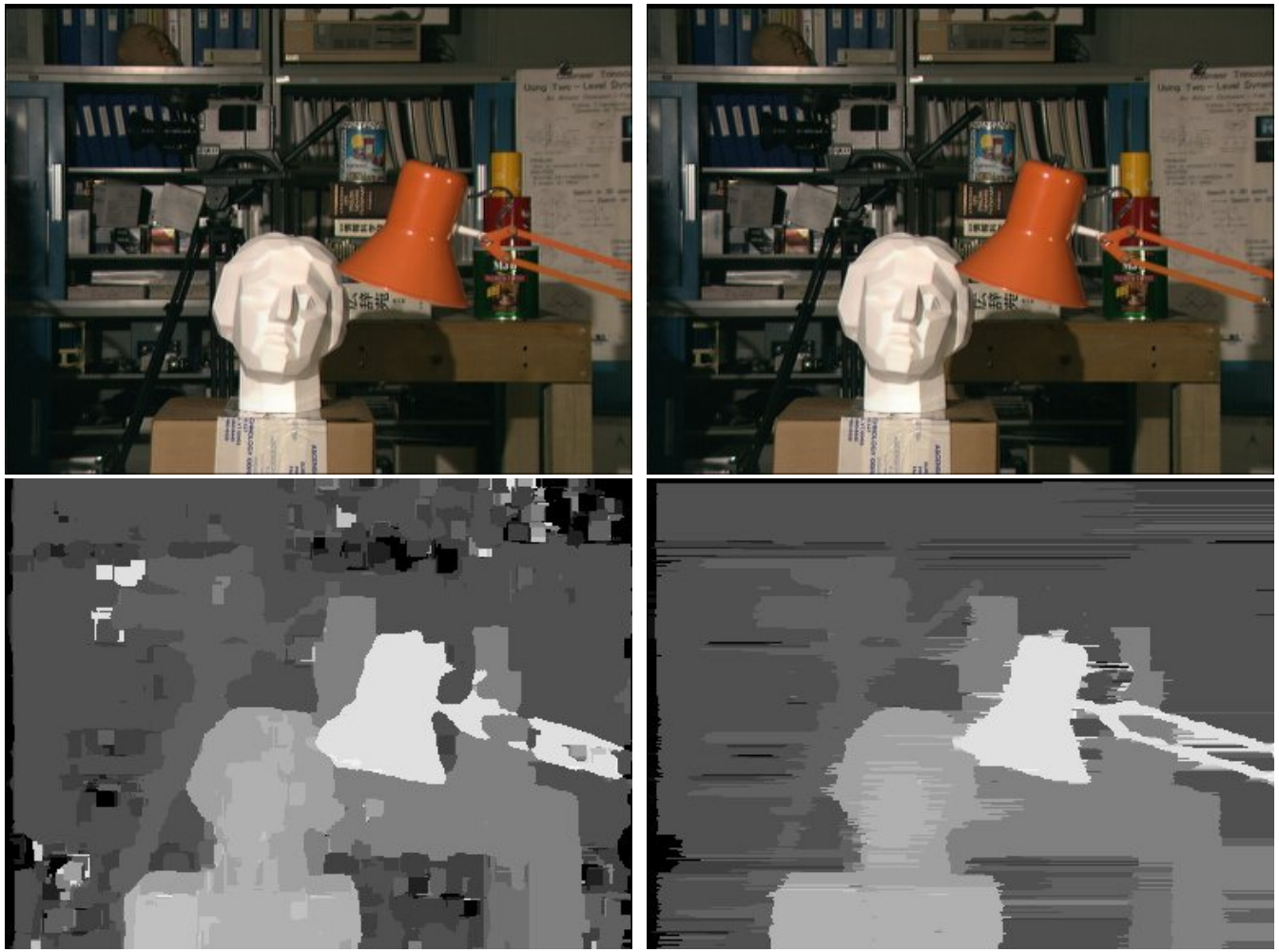
Ignore matches that are either behind camera or too close.

Optimize per scanline: Dynamic Programming



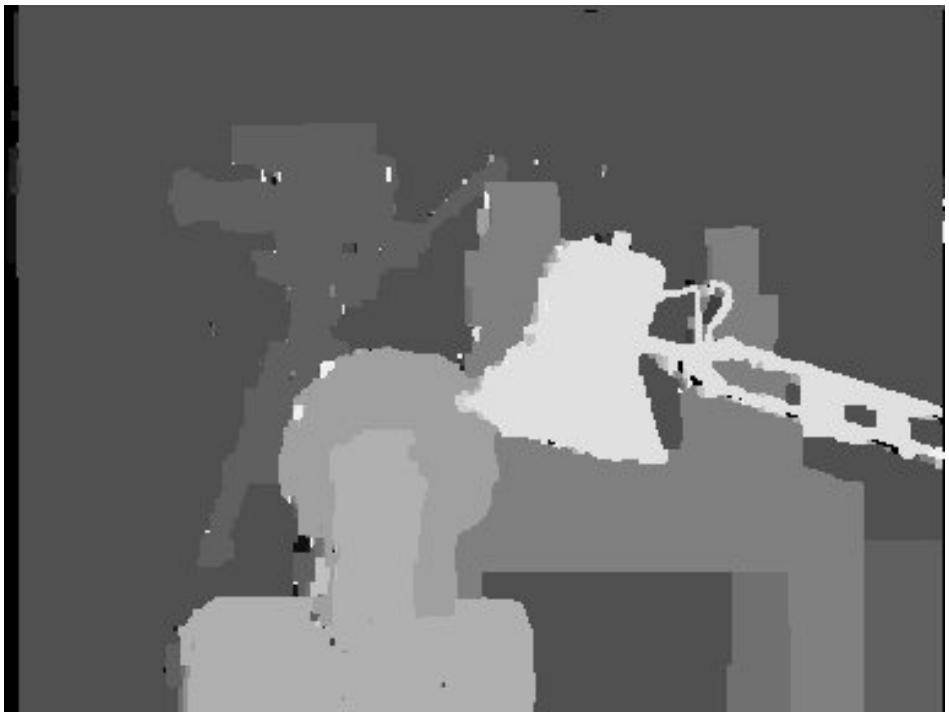
- Set up an energy minimizing problem to optimize per scanline.
- Find lowest cost path using Dynamic Programming (Start to End).
- From one match you can move in three directions
  - Diagonally, stay on same disparirity (low cost).
  - Horizontally or vertically, occlusion in either camera (high cost).

Stereo results



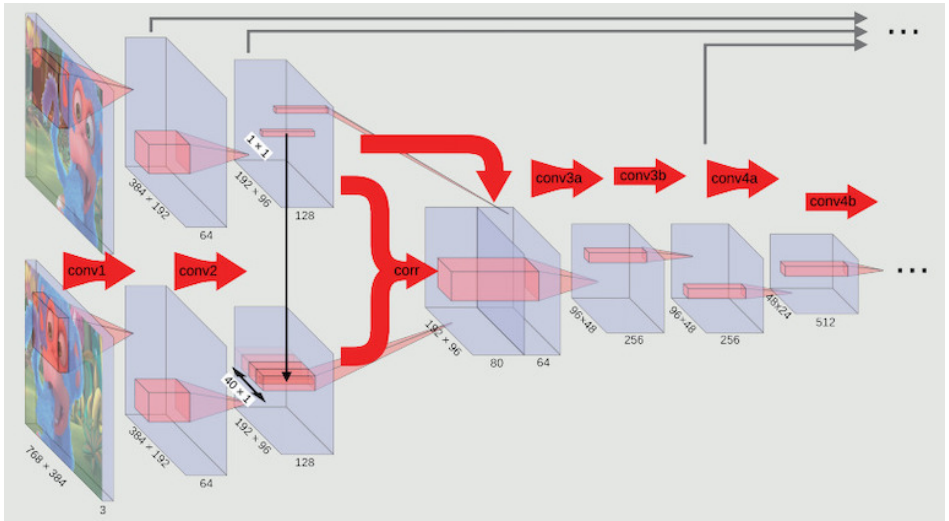
Results based on SSD (left) and dynamic programming (right).

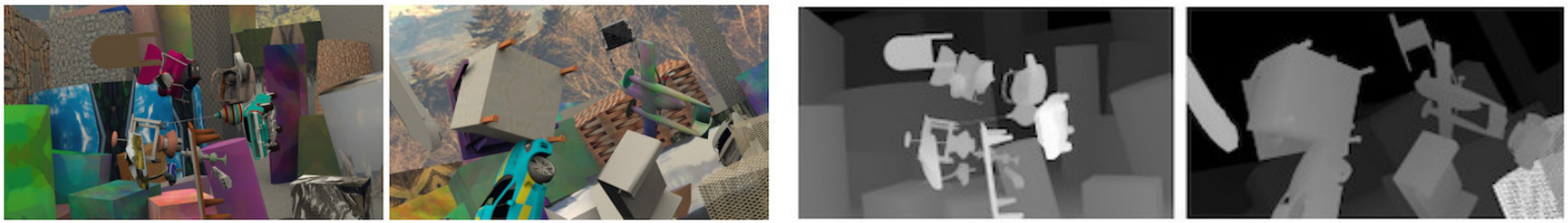
Stereo results with energy minimization



- Better results with global optimization (every pixel simultaneously).
- Can be done through e.g. energy minimization with graph cuts.

Stereo results with deep learning





- Most deep learning methods (e.g. DispNet) also use correlation.
- First a network to extract features, followed by correlation and another network to interpret the correlation scores.

Mayer et al., “A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation”, CVPR, 2016.

NeRF: Neural Radiance Fields (UC Berkeley)

Figure: Models (colour, density) as a function of (position, direction).

NeRF: Neural Radiance Fields

Figure: From a couple of mobile phone images matched with `colmap` and 4 hours of training.

Summary of good questions

- How does stereo work in general?
- Why can you get double vision?
- What is gaze direction of a cyclopean eye?
- What is the relationship between disparities and depths?
- Why does the error in depth increase for larger distances?
- What are the key concepts of epipolar geometry?
- What is an essential matrix and how is it used?
- What might complicate stereo matching?
- What is a disparity space image and why is it useful?

Readings

- Szeliski: Chapters 12.1 – 12.5