Vision and Image Processing: Stereo, Stitching

Søren Olsen

Department of Computer Science
University of Copenhagen

Plan for today

- Stereo correspondance analysis
- Epipolar line geometry, Fundamental matrix
- Triangulation
- Coare-to-fine analysis
- Stitching

What can we see with two eyes

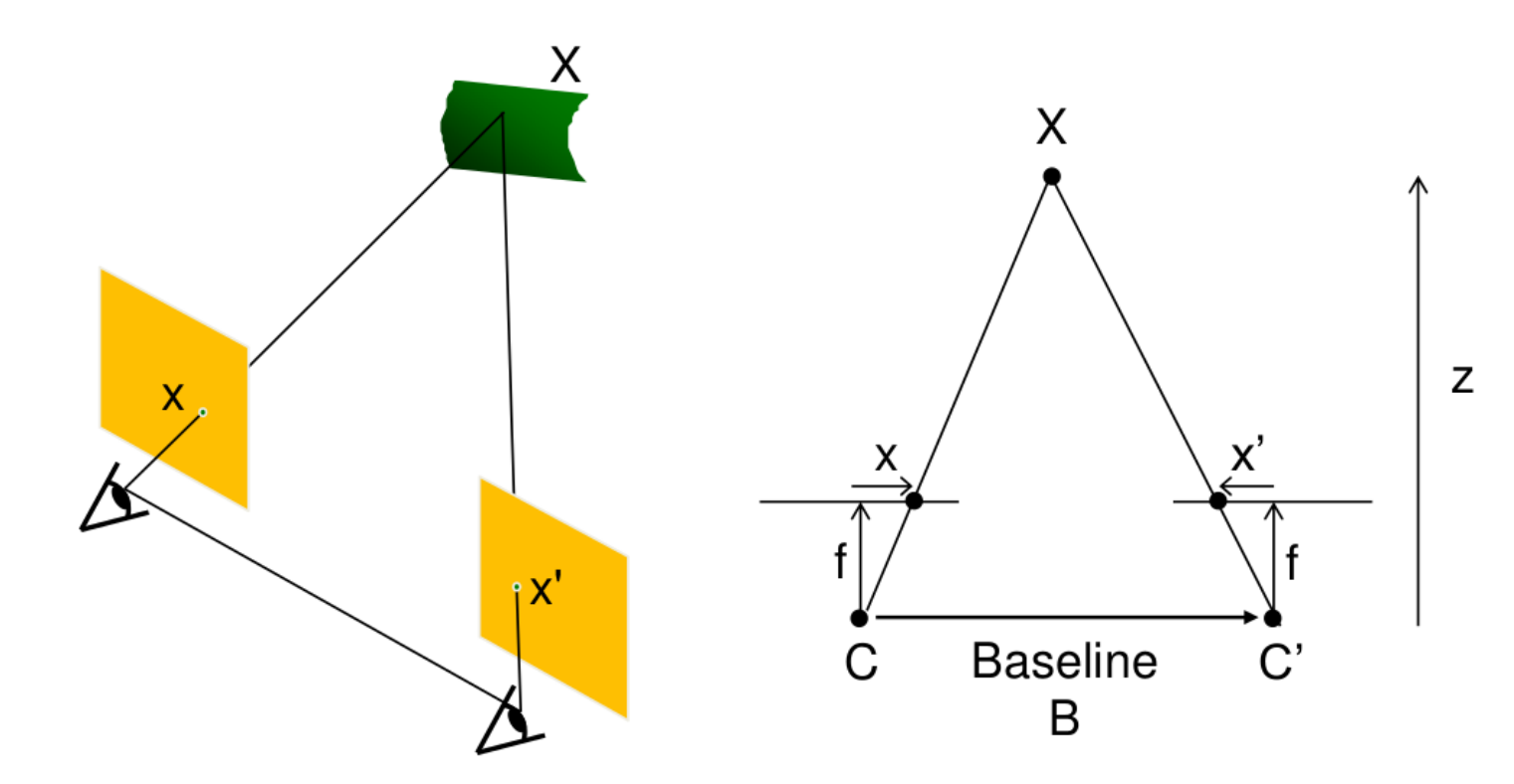
Stereo vision is among the most important human sensing methods.
Next lecture we will talk a lot about stereo vision.

Stereo

In stereo we analyze two images of the same scene but obtained from different viewpoints. Physical scene points project to corresponding image points.

- How are corresponding points related ?
- May we reduce the 2-dimensional correspondence problem to 1D ?
- Can we estimate relationship and is it numerically stable ?

Multiple View Correspondences



If we can recover x' from x we can recover depth: $z = -\frac{fB}{x'-x}$.

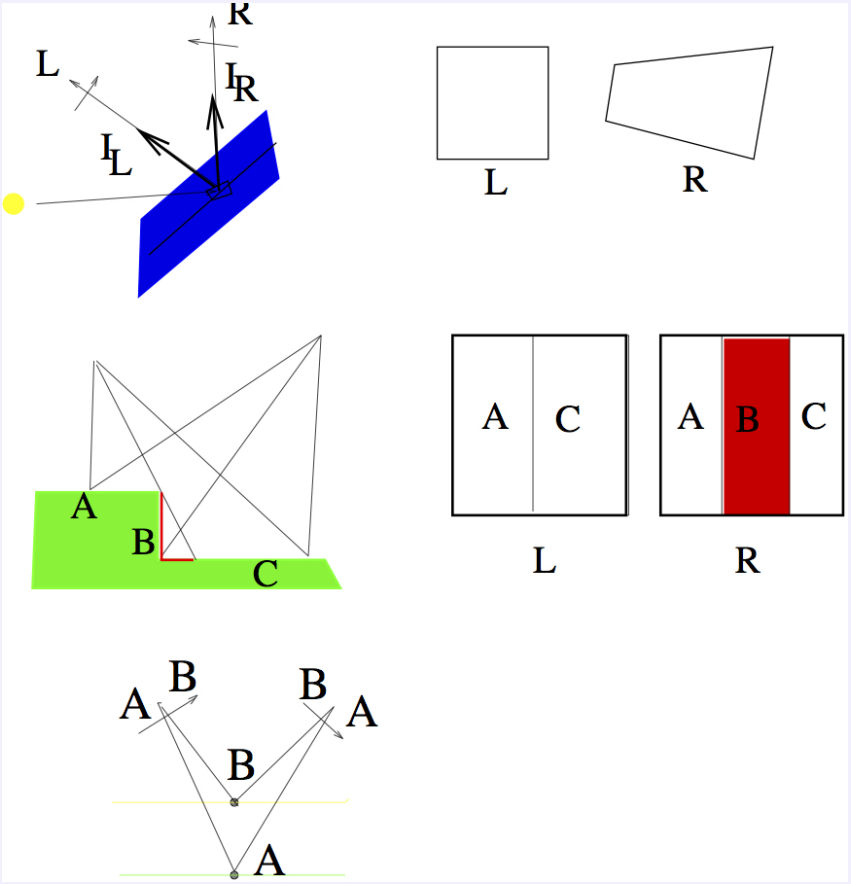
Correspondence analysis

Problem statement: Establish pairs $(\mathbf{p_L}, \mathbf{p_R})$ of image points $\mathbf{p_L}$ in the left image and $\mathbf{p_R}$ in the right image such that both points are projections of the same physical scene point.

- Correspondence analysis is the difficult part of stereo analysis
- Correspondence analysis is the basic of many other applications, eg. stitching, geo-referencing, image alignment/warping etc.
- Most mammalians have stereo vision
- Except for auto-focus cameras, stereo is the most widely applied passive technique for 3D-measurement.

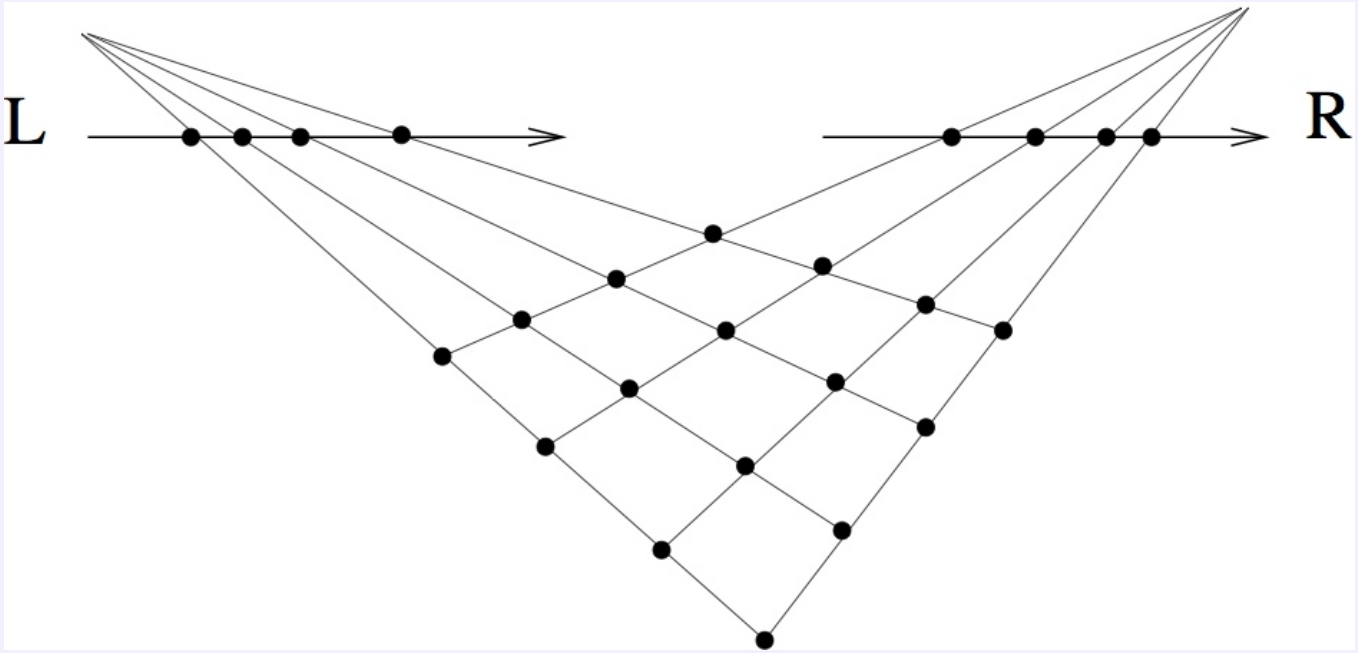
Problems

- Intensity in corresponding points are not equal: $E_L \neq E_R$.
- Many geometric properties, eg. orientation, are not preserved under perspective projections.
- Occlusions: Things/areas visible en one image is invisible in the other.
- Double nail illusion



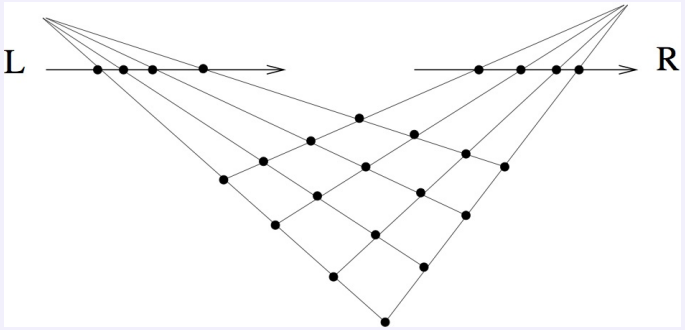
More problems

- Lack of texture/structure/intensity variation makes makes feature matching difficult and intensity comparison vulnerable.
- Complexity of correspondence problem: Given point in one image, how many possible matches exist in the other image ?



Complexity

Assume n points along both epipolar lines. $N(n)$ is number of solutions.



Without assumptions: $N(n) = 2^{n^2}$. $N(4) = 65536$

Assume that each point may match at most one other point:
$N(n) = \sum_{i=0}^{n} \frac{(n!)^2}{(n-i)!(i!)^2}$. $N(4) = 204$.

Assume ordering, ie. $x_L^1 \leq x_L^2 \Rightarrow x_R^1 \leq x_R^2$, and uniqueness:
$N(n) = \frac{(2n)!}{(n!)^2}$. $N(4) = 70$.

Assume all L-points match exactly one R-point:
$N(n) = n!$. $N(4) = 12$.

Assume strong ordering and uniqueness:
$N(n) = 1$.

Simplifying assumptions

- Intensities are similar, eg. $|E_L - E_R| \leq \theta$ or are spatially correlated (more later).
- Fundamental matrix is estimated $\Longrightarrow$ matching is reduced to 1D along epipolar lines.
- The world consist of solid textured surfaces. Thus, the disparity is a single-valued function, and there exist a *unique* solution to the correspondence problem.
- Occlusions and depth discontinuities do not exist.
- Ordering: Corresponding points appear in the same order along the epipolar lines.
- The magnitude of the disparity gradient is limited (for humans to about 1).

Example: Large disparity gradient 1

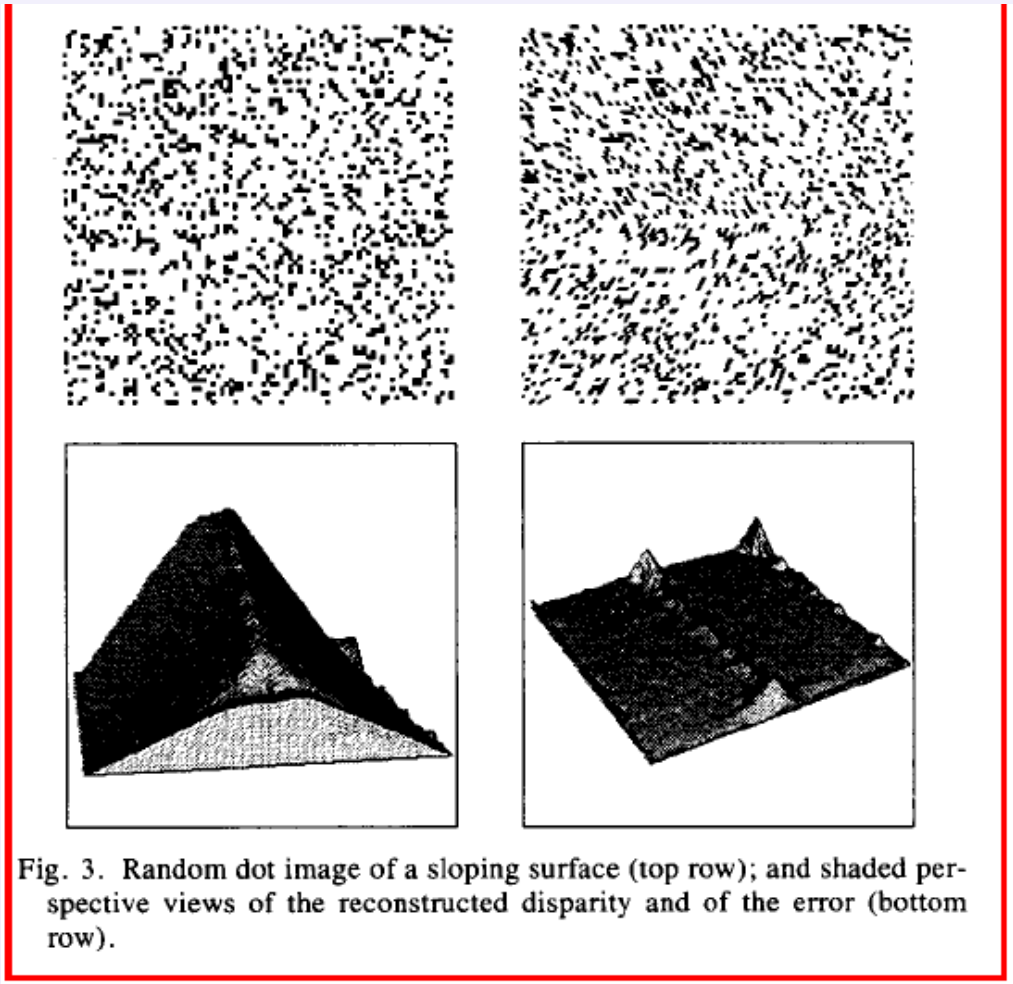


Fig. 3. Random dot image of a sloping surface (top row); and shaded perspective views of the reconstructed disparity and of the error (bottom row).

Humans can fuse random dot stereograms with no structural information. Image has a disparity gradient of 0.6. Humans cannot fuse images with gradient larger than 1.

Example: Large disparity gradient 2

Bidonville on hillside south of Gingerbread District. *(I find seeing this in 3D helpful because one can see how steep the slopes are, as mini-landslides are responsible for a lot of the destruction in the squatter settlements. The destruction does not to be so pervasive in this settlement despite the steep grades.)*

Example: What surface ?

Correspondence analysis

- Dense intensity based methods may be accurate but is very noise sensitive and have a small capture area. Also, they may be computationally expensive.
- Sparse feature based methods is faster, more reliable, and have larger capture area, but results in scattered depth information.
- Very local features as edge points often has a short descriptor, e.g. edge orientation.
- Less local features (as SIFT) is less dense, but often has a more expressive descriptor.
- Large features (as image segments) are few and more easy to match, but gives less depth information.

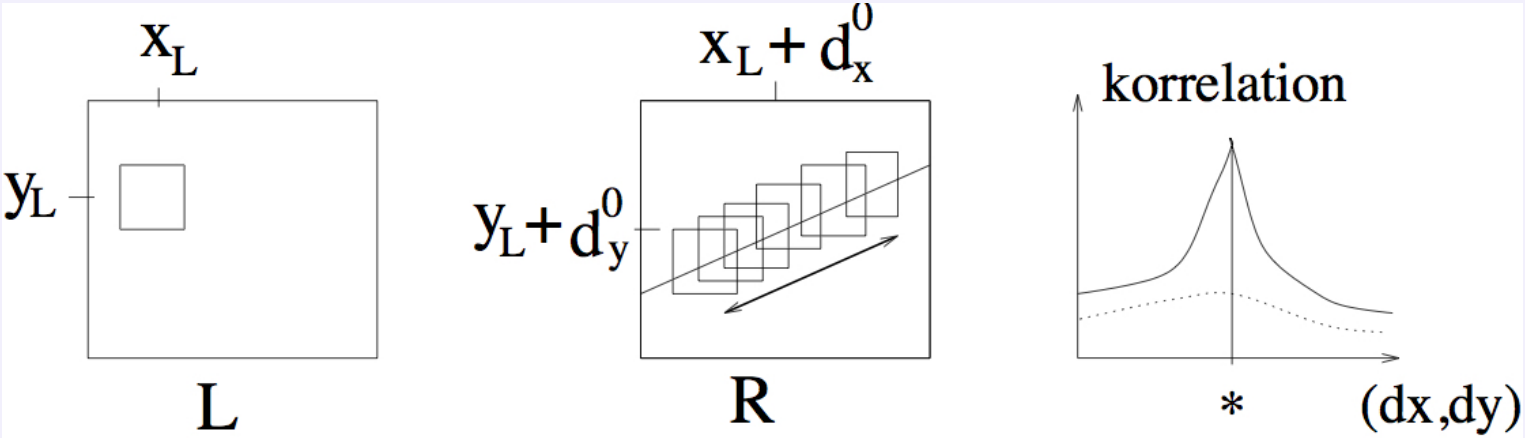
Feature based stereo

- Detect interest (eg. SIFT) points in both images
- Attribute interest point with (eg. SIFT) descriptors
- Match the points with most similar descriptors
- Eliminate false matches that does not comply with the fundamental matrix equation
- Gather more matches or fit dense surface (optional)

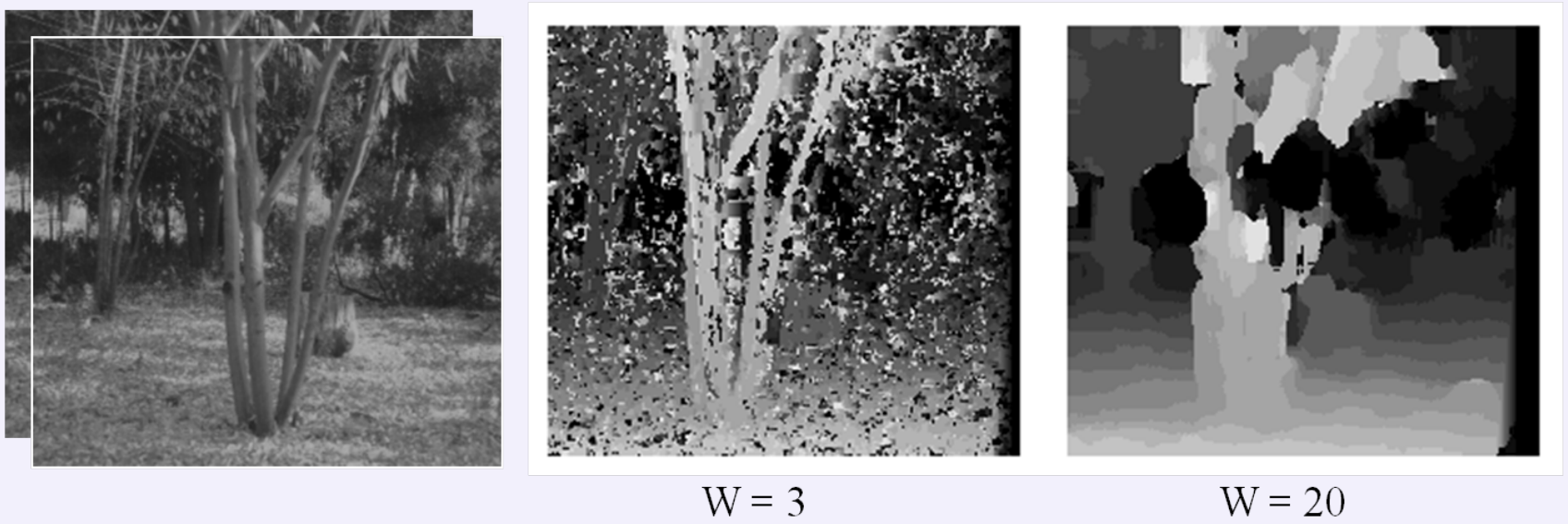
Procedure above often done coarse-to-fine to speed up and reduce blunders

Area based stereo

- Pixelwise intensity comparison does not work. Areas, say 7×7, or 13×13 are used. Larger windows implies better robustness, less precision and larger vulnerability to occlusions.
- All R-windows centered and displaced along the epipolar line are compared to the L-window centered at the point in question and the best is chosen.
- Typical measure: Normalised cross-correlation



Disparity Map By Dense Block Matching[1]



- Window size 3: Noisy but detailed.
- Window size 20: smoother, but missing details.

[1] Slide adapted from Derek Hoiem

Cross-correlation

The cross-correlation between two continuous functions (with limited square integral) is defined by:

$$h(x) = (f \circ g)(x) \quad = \quad \int f^{\star}(\alpha) g(x \; + \; \alpha) d\alpha$$

Discrete normalised 2D cross correlation is defined by:

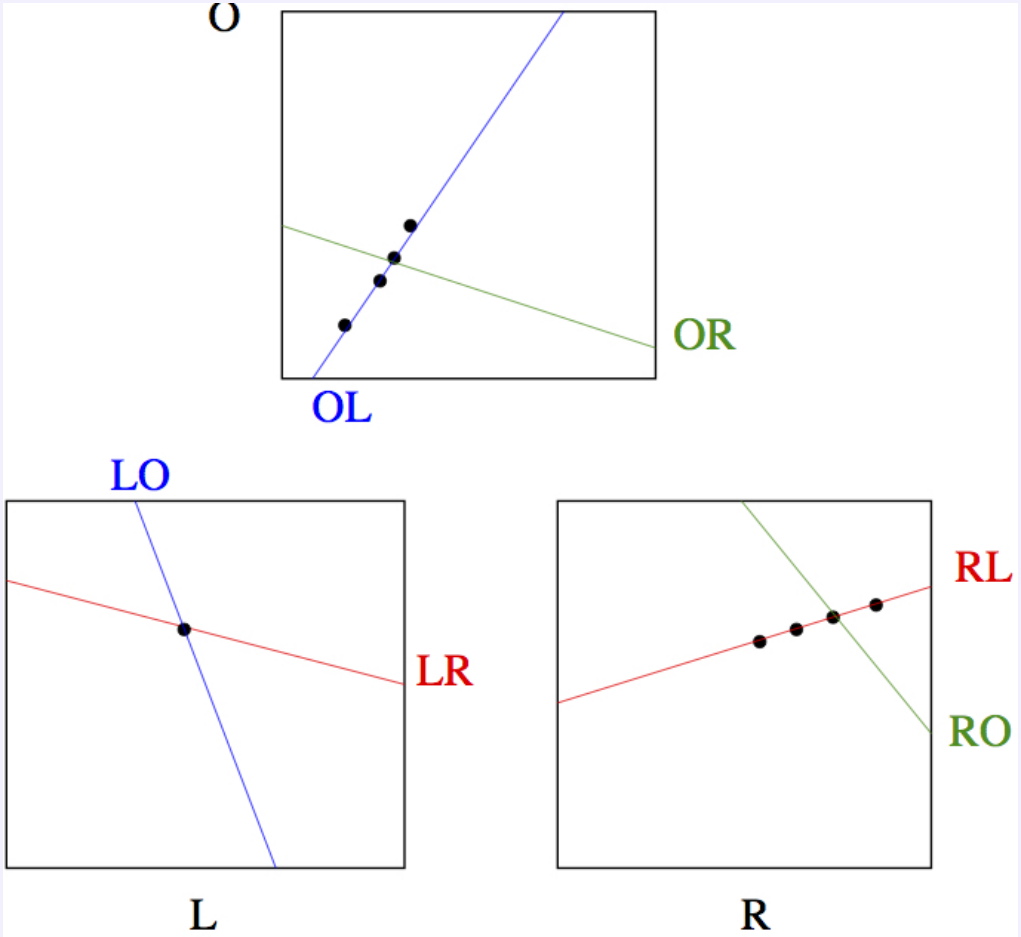
$$\frac{1}{n} \sum_{(x,y) \in \Omega} \frac{(f(x,y) - \bar{f}) \cdot (g(x+\alpha, y+\beta) - \bar{g})}{\sigma_f \sigma_g}$$

where n is the number of pixels in Ω, $\bar{f}$ is the mean value of f in Ω, σ_f is the standard deviation of f within Ω (and similarly for g).

Cross-correlation is used in Template matching where we are searching for positions in $f(x, y)$ where the signal/image is identical/similar to the prototype $g(x, y)$. Such positions can be identified as the local maxima's of $(f \circ g)(x, y)$.

3-Camera stereo

The use of 3 cameras in stereo vision, and assuming all fundamental matrices known, makes the correspondence analysis more easy and robust.



Scale-Space -repetition

Often we don't know the size of the things we are imaging, so we have to use both large and small filters when we analyse images. In practice we represent each image at a number of scales.

A scale-space representation is obtained by convolving the image with Gaussian kernels of increasingly larger size. At small scales we only blur the images slightly to attenuate noise. At large scale we blur heavily to ensure that only large scale structures survive.

To save space and (in particular) time we subsample the smoothed image versions. The result is an image pyramid.

Coarse-to-fine

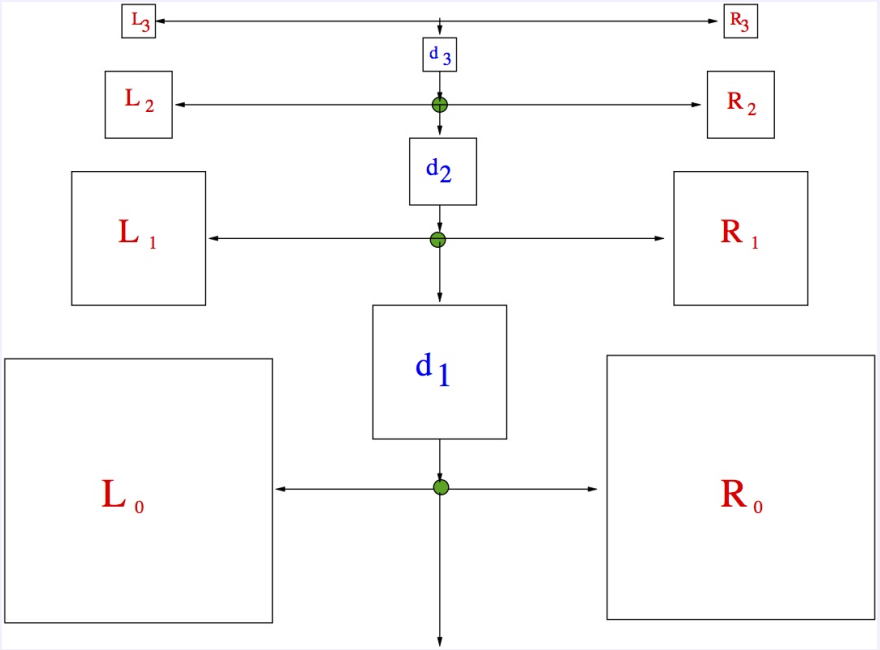
Pyramid-based coarse to fine approaches:

- reduce the time complexity from $\mathcal{O}(N)$ to $\mathcal{O}(\log(N))$.
- reduce the complexity of the correspondence problem with a large factor (see previous slide).
- Facilitates global operations using only local computations

Principle: Use approximate solutions obtained at higher pyramid level to constrain the search at lower levels.

Coarse-to-fine Stereo

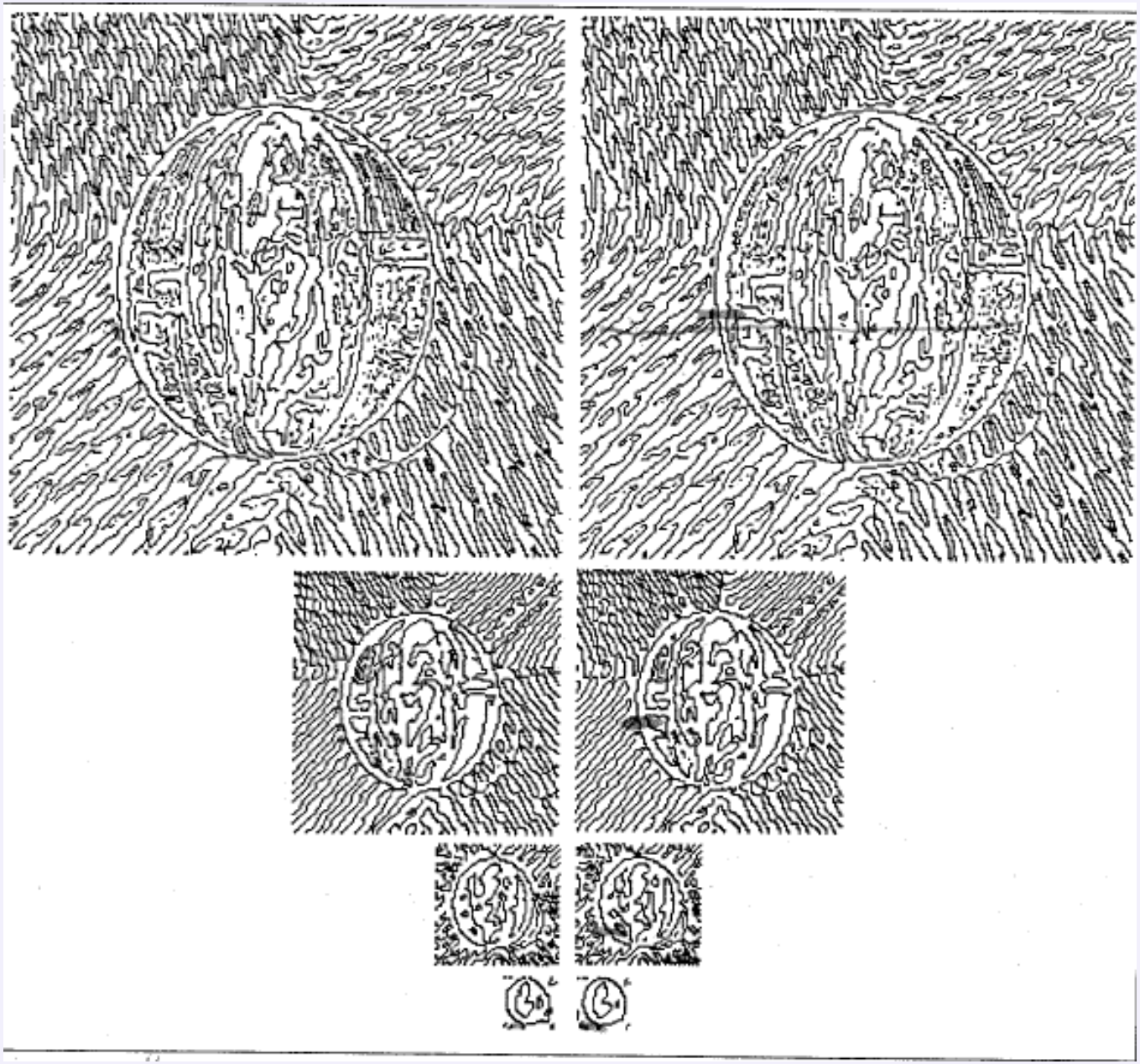
In practice the disparity may be large (several hundred pixels) and have large variation (eg. from -50 to +50 pixels). To reduce the size of the search area we need an estimate of the disparity.



Method: Succesive smoothing and downsampling

The total space requirement is:

$$1 + \frac{1}{4} + \frac{1}{16} + \cdots < \frac{4}{3}$$



Tsukuba

The images in this and the next slides are from Scharstein, Szeliski: *A taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms*, Int.Jour. of Comput.Vis. 47, 2002. See: `http://vision.middlebury.edu/stereo/data/`.

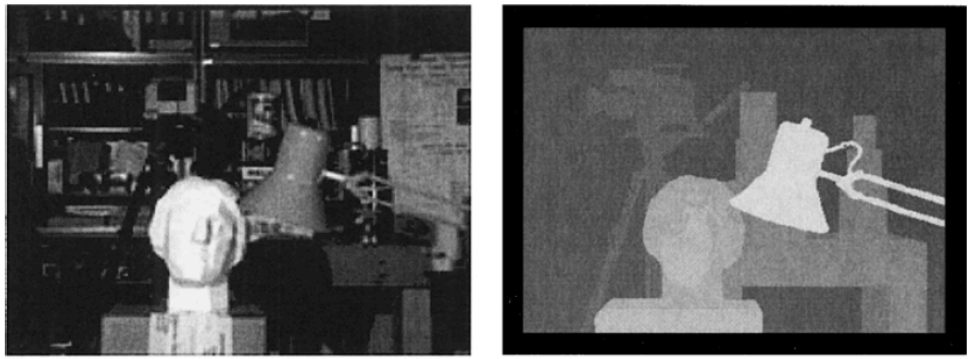
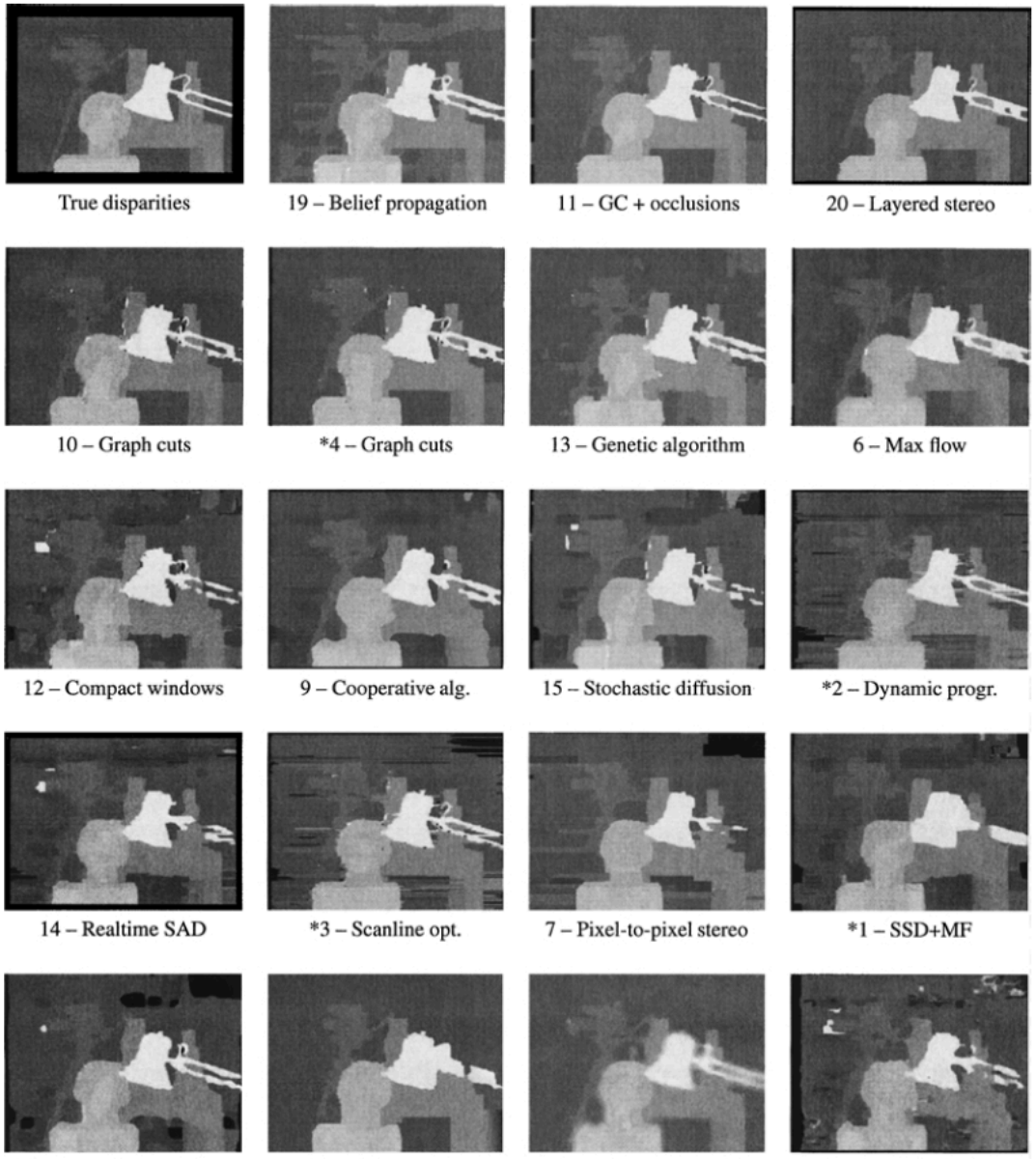


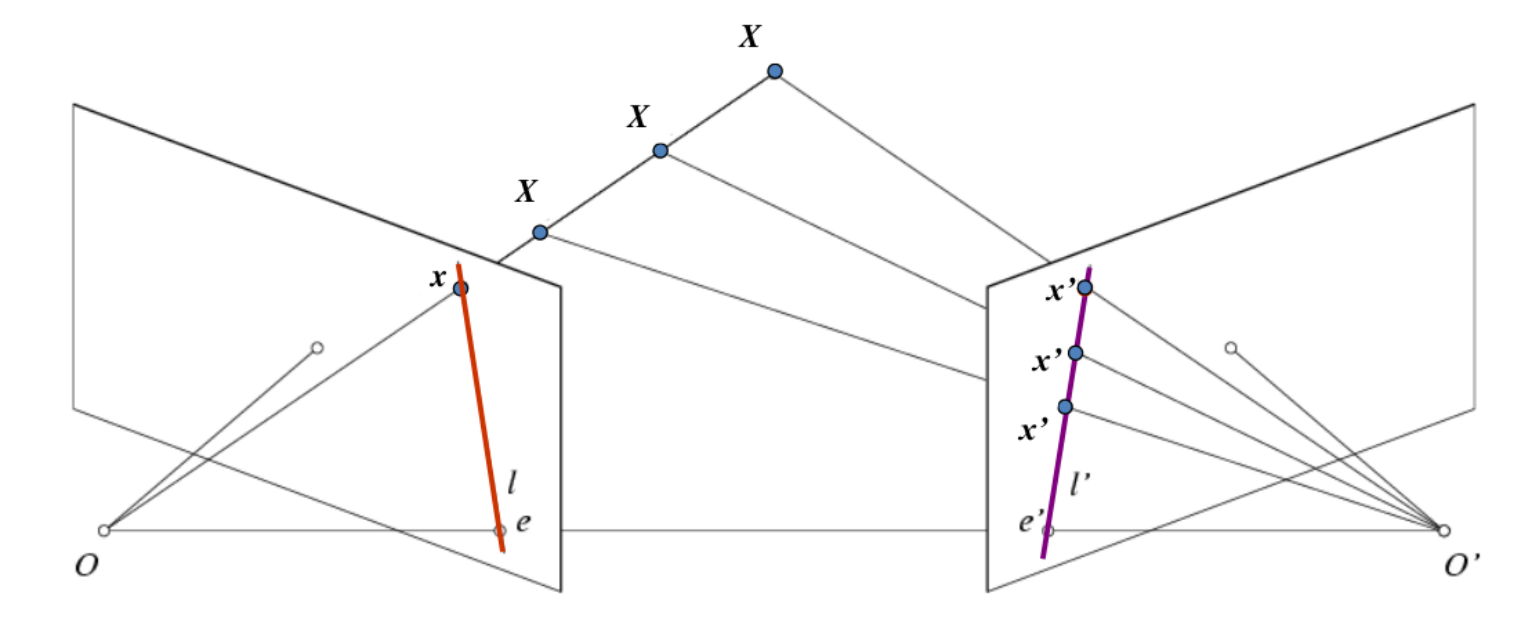
Figure 4. Segmented region maps: (a) original image, (b) true disparities, (c) textureless regions (white) and occluded regions (black), (d) depth discontinuity regions (white) and occluded regions (black).

Other peoples results



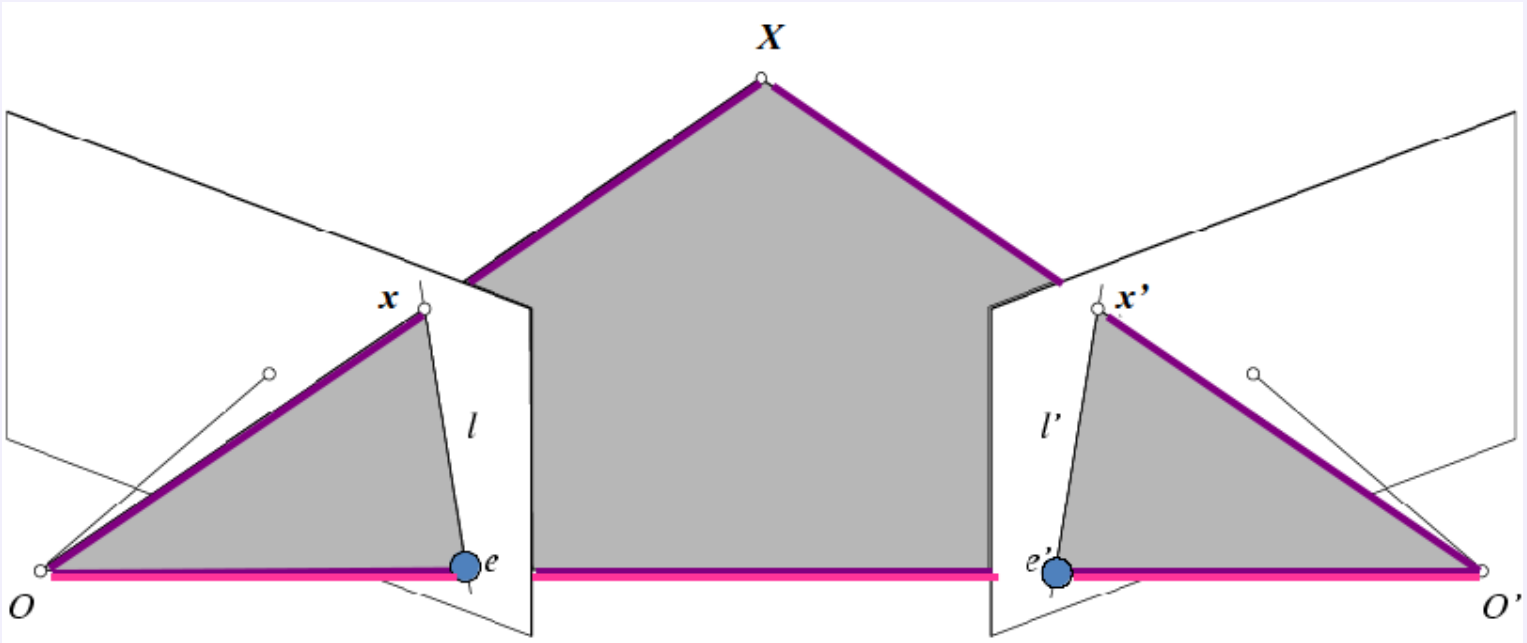
QUESTIONS ?

Epipolar Constraints



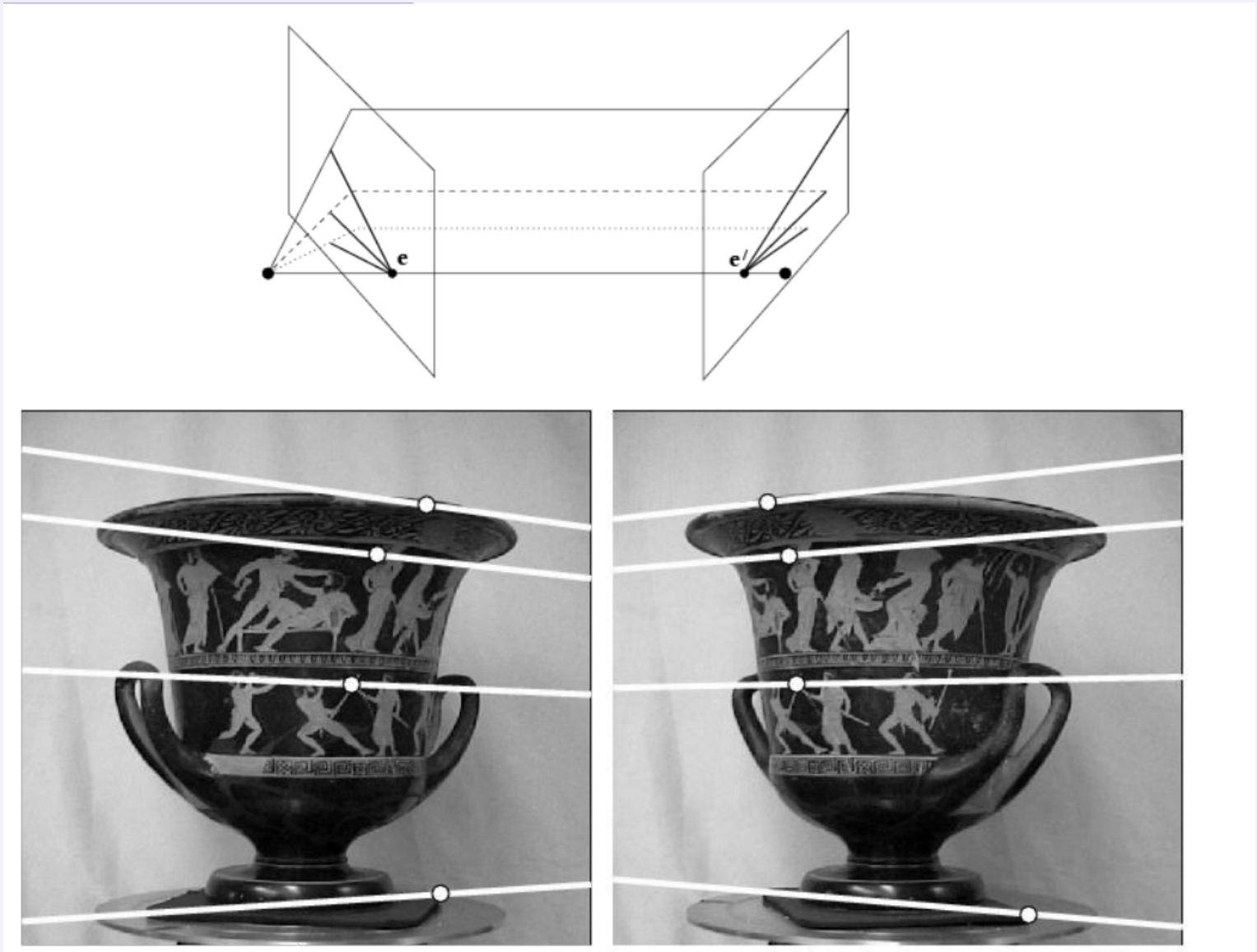
- Corresponding point for x must lie in corresponding line l'
- Corresponding point for x' must lie in corresponding line l

Epipolar Constraints

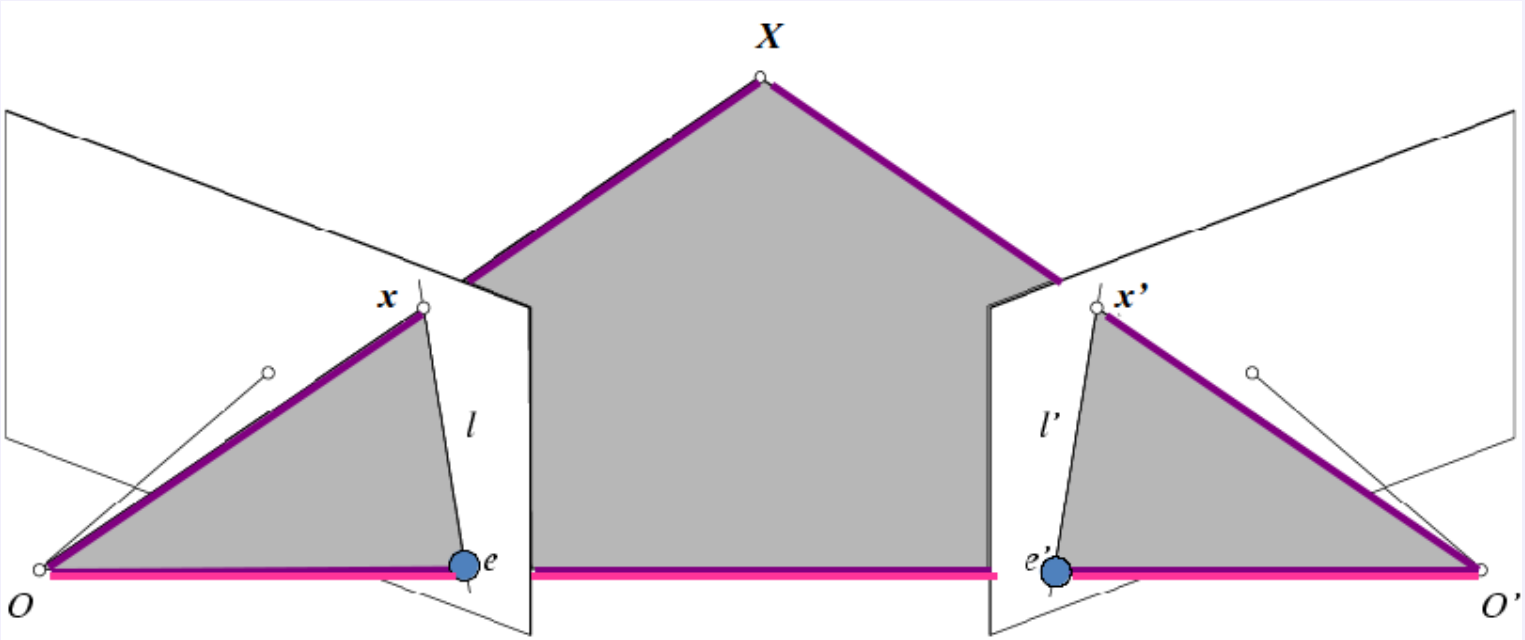


- Line connecting O and O': baseline
- Plane through baseline and x or x': Epipolar Plane
- Epipoles: intersection of baseline and image planes: projection of the other camera center.
- Epipolar Lines - intersections of epipolar plane with image planes (always come in corresponding pairs)

Example: Converging cameras



Epipolar Constraints



- Line connecting O and O': baseline
- Plane through baseline and x or x': Epipolar Plane
- Epipoles: intersection of baseline and image planes: projection of the other camera center.
- Epipolar Lines - intersections of epipolar plane with image planes (always come in corresponding pairs)

The essential matrix E

- Let y and y' be 3D coordinates of the same scene point in the two (different) 3D-**camera** coordinate systems. The two systems are related by a rotation and a translation.

$$y' = R(y - \mathbf{t})$$

- We will later show that corresponding points y_c and y'_c are related by a 3×3 matrix E built from R and t

$$y_c^T E y'_c = 0.$$

- E is called the essential matrix (Longuet-Higgins 1981).

The fundamental matrix F

- For corresponding points, the image coordinates (x and x') are related to the camera coordinates (y and y') through the calibration matrices K and K'. Thus:

$$0 = y^T E y' = (K^{-1}x)^T E (K'^{-1}x') = x^T (K^{-T} E K'^{-1}) x' = x^T F x'$$

  where x and x' are the homogene representation of the corresponding points in image coordinates, and where F is the fundamental matrix.

- Given a sufficient number of corresponding image points (x and x'), the fundamental matrix F can be estimated.
- Given the internal intrinsic parameters (K and K') the essential matrix E may be computed from F.
- Given E, the position and orientation of camera 1 vs camera 2 (i.e., R and $\mathbf{t}$) can be recovered.

The Fundamental matrix

Ther fundamental matrix is a 3×3 matrix that relates corresponding points in a bilinear homogeneous equation:

$$[x_L,\ y_L,\ 1]\ F \begin{bmatrix} x_R \\ y_R \\ 1 \end{bmatrix} = 0$$

F has 9 elements, but it can be shown that $\det F = 0$, so it has only 7 degrees of freedom (independent parameters).

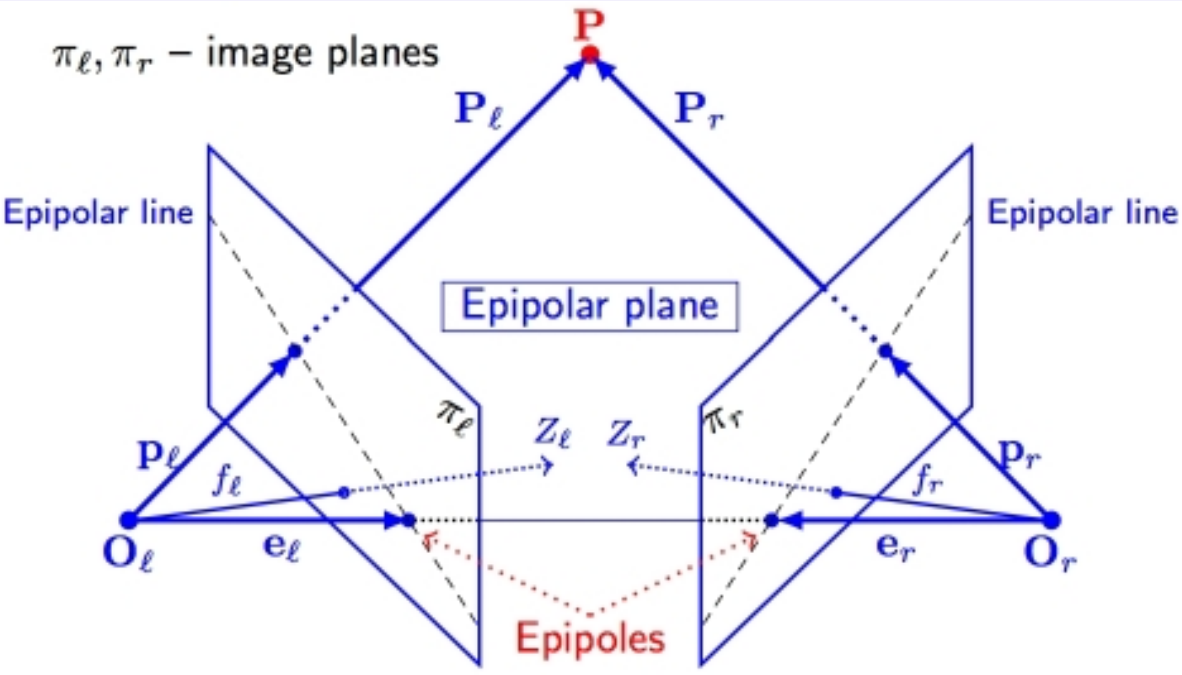
Fixing either the left or right image point gives a straight line (the epipolar line) in the other image.

Calibration and reconstruction

- Given (sufficient) image point correspondences, the fundamental matrix F may be estimated using linear algebra.
- Linear estimation of F is easy, but not accurate. In practice a non-linear post- non-linear optimisation is needed.
- Given F, the stereo correspondence problem is reduced to a one-dimensional search along the epipolar lines.
- Given F, and the internal camera parmeters$K$ and K', then the reconstructions of 3D points is possible (using linear algebra) from image point correspondences.

If you want to know more, you must take the ATIA-course.

Proof of $x_R^\top E x_L = 0$



We have that the camera coordinate systems are related by:

$$\mathbf{P}_R = R(\mathbf{P}_L - \mathbf{T})$$

Definition

The coplanarity condition: $\mathbf{P}_L$, $\mathbf{T}$, and $\mathbf{P}_L - \mathbf{T}$ are all in the epipolar plane. Then, also $R^\top \mathbf{P}_R$ is within the plane.

The cross-product

The cross-product between two vectors **a** and **b** is a vector that is perpendicular to both:

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} -a_3 b_2 + a_2 b_3 \\ a_3 b_1 - a_1 b_3 \\ -a_2 b_1 + a_1 b_2 \end{pmatrix} = S\mathbf{b}$$

where

$$S = [a]_x = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$$

We see that S is an anti-symmetric and rank deficient matrix. S has rank 2.

Proof cont. 2

Because $\mathbf{P}_L$, $\mathbf{T}$, and $\mathbf{P}_L - \mathbf{T}$ all are in the epipolar plane we can write:

$$\begin{aligned} 0 &= (\mathbf{P}_L - \mathbf{T})^\top \mathbf{T} \times \mathbf{P}_L \\ &= (R^\top \mathbf{P}_R)^\top \mathbf{T} \times \mathbf{P}_L \\ &= (R^\top \mathbf{P}_R)^\top S \mathbf{P}_L \\ &= \mathbf{P}_R^\top R S \mathbf{P}_L \\ &= \mathbf{P}_R^\top E \mathbf{P}_L \end{aligned}$$

where we have used that $\mathbf{P}_R = R(\mathbf{P}_L - \mathbf{T})$ and $E = RS$.
Since $\text{rank}(S) = 2$, $\text{rank}(E) = 2$.

The fundamental matrix equation once more

We have now established the Essential matrix equation $\mathbf{P}_R^\top E \mathbf{P}_L = 0$. To get to the fundamental matrix equation we remember the relation between the camera- and the image coordinate systems:

$$K = \begin{pmatrix} \alpha & s & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

Using $\mathbf{p}_L = K_L \mathbf{P}_L$ and $\mathbf{p}_R = K_R \mathbf{P}_R$ and defining

$$F = K_R^{-\top} E K_L^{-1}$$

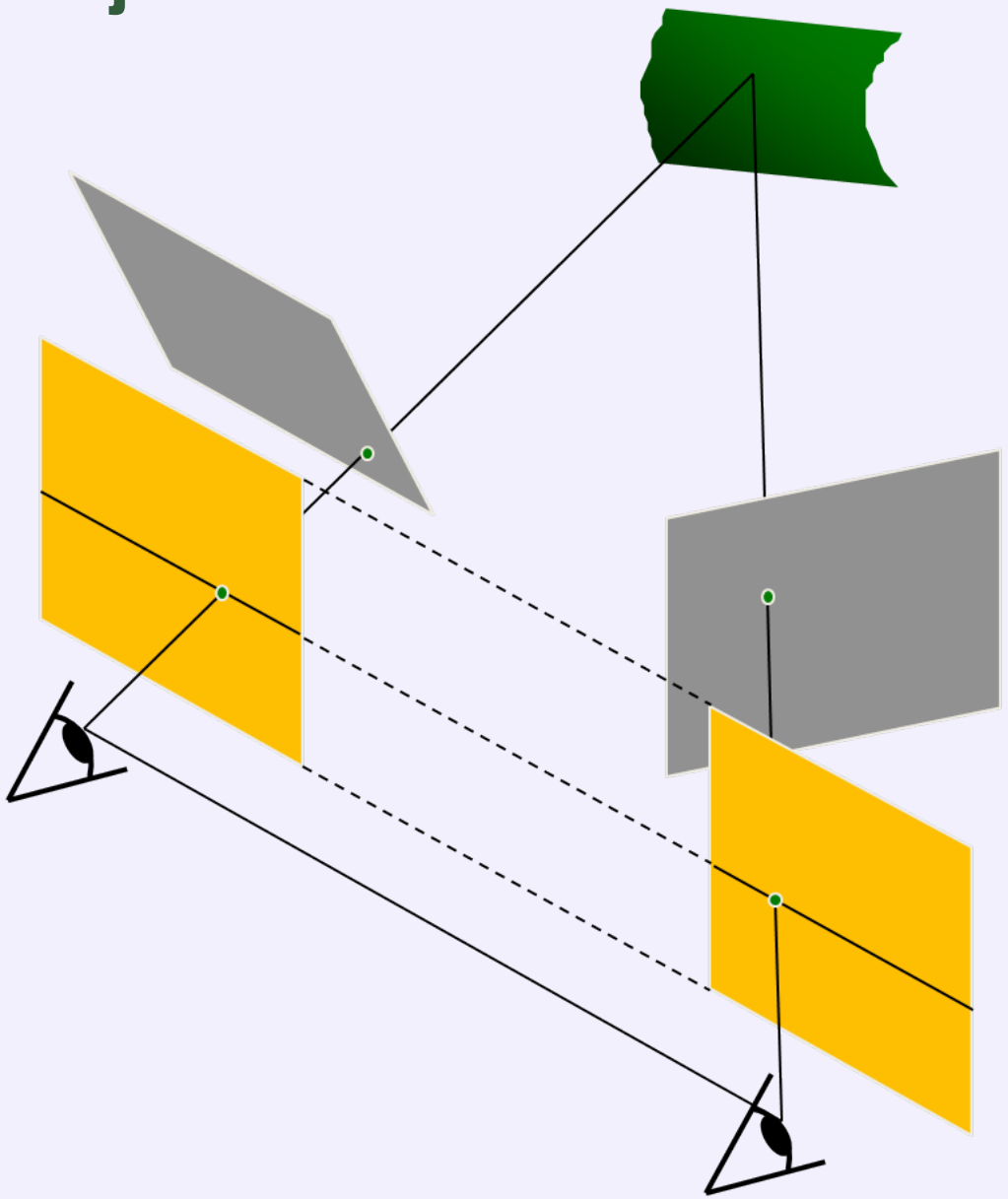
we finally get:

$$\mathbf{p}_R^\top F \mathbf{p}_L = 0$$

Non horizontal Scan lines

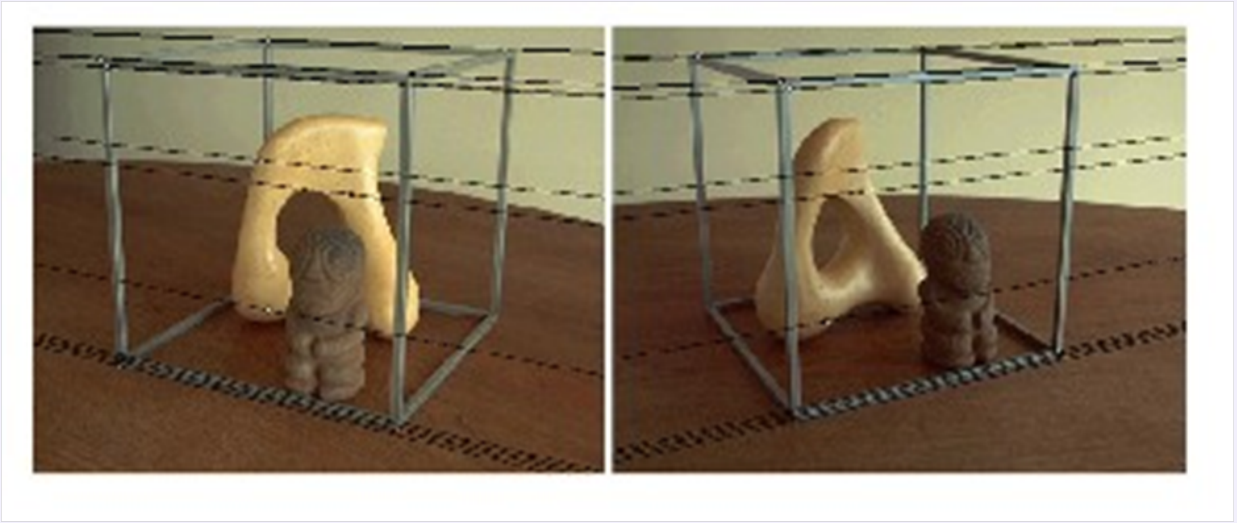
- If calibration known, the essential matrix provides epipolar constraints.
- What when cameras are in general position and calibration is unknown?
- Non calibrated views: Estimate the fundamental matrix.
- Knowing Essential or Fundamental matrix allows (almost) for image rectification.

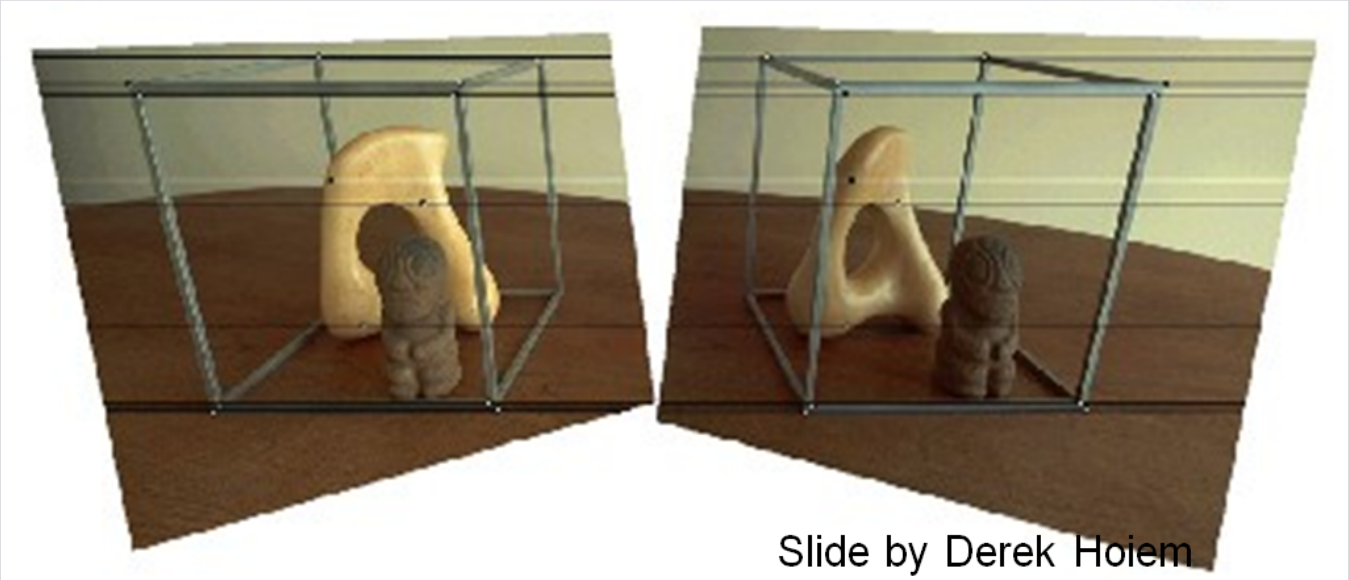
Projective Rectification



- Reproject onto a common plane parallel to line between camera centers
- Projections are homographies!
- Pixel motion is horizontal after reprojection.
- Cf Loop-Zhang, CVPR 1999 (Rectification is not easy)

Projective Rectification example





Remember the Camera Matrix

- Camera calibration matrix, now extended with Image plane transformation (axis scalings, shear, translation)

$$\mathbf{C} = \mathbf{K}\,[\mathbf{R}\ \mathbf{t}]$$

- $\mathbf{K}$ 3×3 matrix encoding the homogeneous transformations inside the camera. $\mathbf{K}$ specifies the Intrinsic parameters.
- $[\mathbf{R}\ \mathbf{t}]$ Concatenation of world coordinates rotation and origin translation to align camera and world coordinates.

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} wx \\ wy \\ w \end{bmatrix} = \underbrace{\begin{pmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{K}} \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{pmatrix}}_{[\mathbf{R}\ \mathbf{t}]} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

Given M, how can we triangulate ?

Let $\mathbf{m}^i$ be the i'th row of the camera matrix and $U = (X, Y, Z, 1)^\top$. Then:

$$w \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{pmatrix} \mathbf{m}^1 \\ \mathbf{m}^2 \\ \mathbf{m}^3 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{m}^1 U \\ \mathbf{m}^2 U \\ \mathbf{m}^3 U \end{pmatrix}$$

We will later se how we may estimate the parameters of the camera matrix and use this to compute 3D structure (depth) from point correspondences.

Linear triangulation for 2 calibrated cameras

Let $U = (X, Y, Z, 1)^\top$. For the first coordinate in the left camera we have:

$$x^L = \frac{\mathbf{m}_L^1 U}{\mathbf{m}_L^3 U}$$

and similar for the y-coordinate and the right camera. Multiplying by the denominator we get:

$$\begin{aligned} \mathbf{m}_L^3 x_L U &= \mathbf{m}_L^1 U \\ \mathbf{m}_L^3 y_L U &= \mathbf{m}_L^2 U \\ \mathbf{m}_R^3 x_R U &= \mathbf{m}_R^1 U \\ \mathbf{m}_R^3 y_R U &= \mathbf{m}_R^2 U \end{aligned}$$

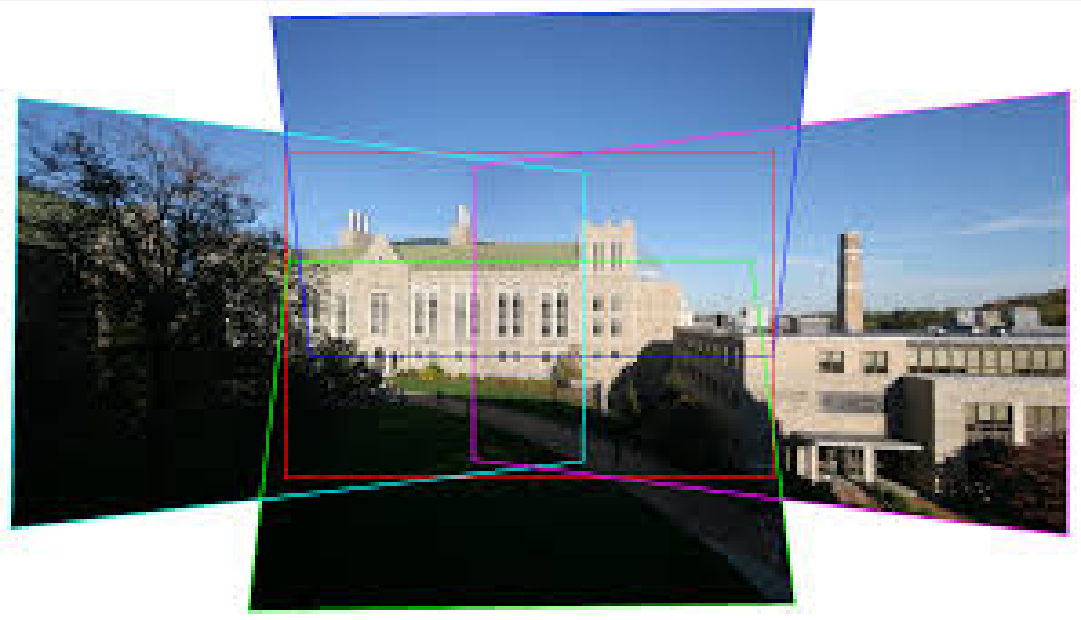
Subtracting the right side and putting U outside a parentesis we get the homogeneous equation $AU = 0$, where:

$$A = \begin{pmatrix} \mathbf{m}_L^3 x_L - \mathbf{m}_L^1 \\ \mathbf{m}_L^3 y_L - \mathbf{m}_L^2 \\ \mathbf{m}_R^3 x_R - \mathbf{m}_R^1 \\ \mathbf{m}_R^3 y_R - \mathbf{m}_R^2 \end{pmatrix}$$

QUESTIONS ?

Stitching

In stitching, or panorama making, several images are combined into one larger image.



A standard transformation technique is to map the images through homographies. However, this requires that the transformed scene surface is planar. Obviously this often is not the case.

The geometric transformation between two images of a planar scene

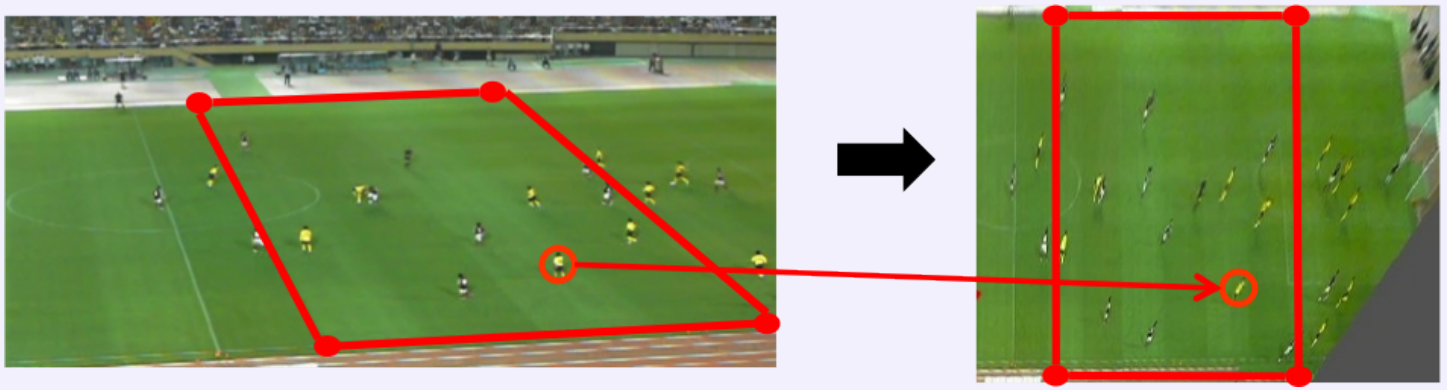
- The perspective projection of a planar scene surface is a transformation $(X, Y, Z) \longrightarrow (x, y)$ called an homography:

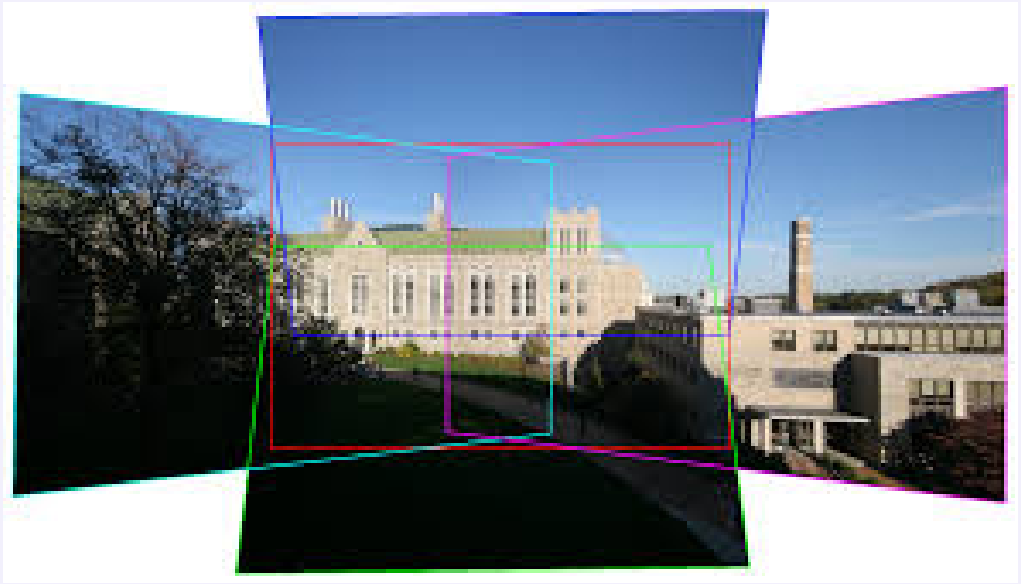
$$\begin{bmatrix} wx \\ wy \\ w \end{bmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = H \cdot \mathbf{X}$$

- The transformation between two images of a planar surface is an homography.
- Homographies conserves straight lines, but not parallelism.
- Two parallel lines intersect at infinity: after an homography they may intersect at finite distance.

Homographies

An homography can accurately describe the stitching of images of a planar scene. Non-planar scenes cannot be stitched correctly using homographies.





Homography estimation 1

Let $\mathbf{x}$ and $\mathbf{x}'$ be corresponding points in homogeneous coordinates.

If $\mathbf{x}' = H\mathbf{x}$ then $\mathbf{x}' \times H\mathbf{x} = \mathbf{0}$ (the vectors have the same direction but may have different magnitude).

Write out the definition of the cross product and isolate the unknown 9 elements. For compactness use $h = (\mathbf{h}_1^\top, \mathbf{h}_2^\top, \mathbf{h}_3^\top)^\top$:

$$\begin{bmatrix} \mathbf{0}^\top & -\mathbf{x}_i^\top & y'\mathbf{x}_i^\top \\ \mathbf{x}_i^\top & \mathbf{0}^\top & -x_i'\mathbf{x}_i^\top \\ -y_i'\mathbf{x}_i^\top & x_i'\mathbf{x}_i^\top & \mathbf{0}^\top \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The last equation may be ignored because it is linearly dependent.

Homography estimation 2

In detail the nine parameters are found as the last column of the matrix V in a SVD-decomposition $A = UDV^\top$ of A, where:

$$A\mathbf{h} = \begin{bmatrix} 0 & 0 & 0 & -x_L^1 & -y_L^1 & -1 & x_L^1 x_R^1 & y_L^1 x_R^1 & x_R^1 \\ x_L^1 & y_L^1 & 1 & 0 & 0 & 0 & -x_L^1 y_R^1 & -y_L^1 y_R^1 & -y_R^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & -x_L^n & -y_L^n & -1 & x_L^n x_R^n & y_L^n x_R^n & x_R^n \\ x_L^n & y_L^n & 1 & 0 & 0 & 0 & -x_L^n y_R^n & -y_L^n y_R^n & -y_R^n \end{bmatrix} \mathbf{h} = \mathbf{0}$$

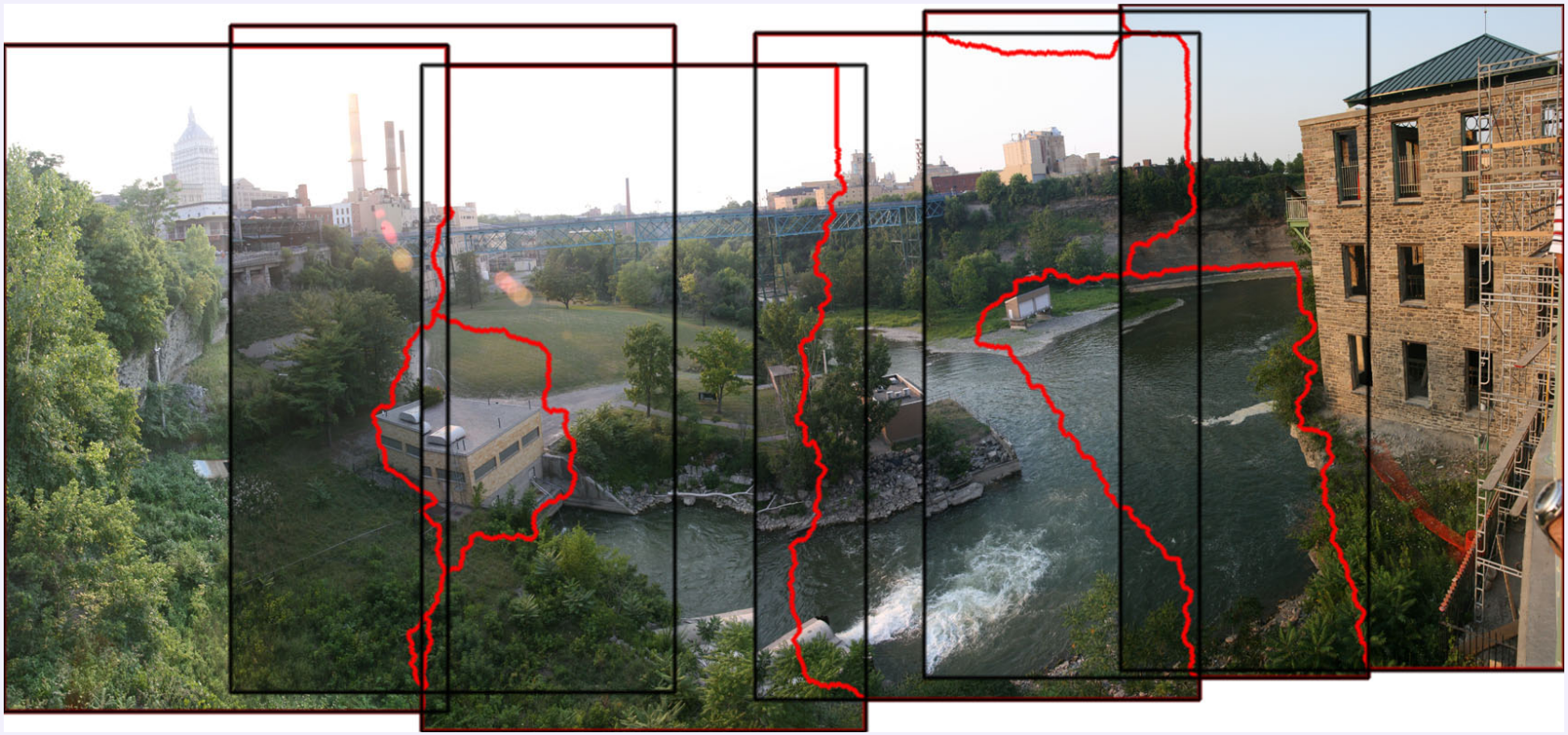
Understanding homographies

Since a homography H can be determined up to scale only we may normalize to make $h_{33} = 1$:

$$\begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{pmatrix}$$

- The two elements in the last colums determines translation
- The two elements in the last row determins perspective distortion
- The four upper left elements determine rotation and scaling

Stitching example with seam-lines



(from Wikipedia)

QUESTIONS ?