Vision and Image Processing: Correspondence analysis

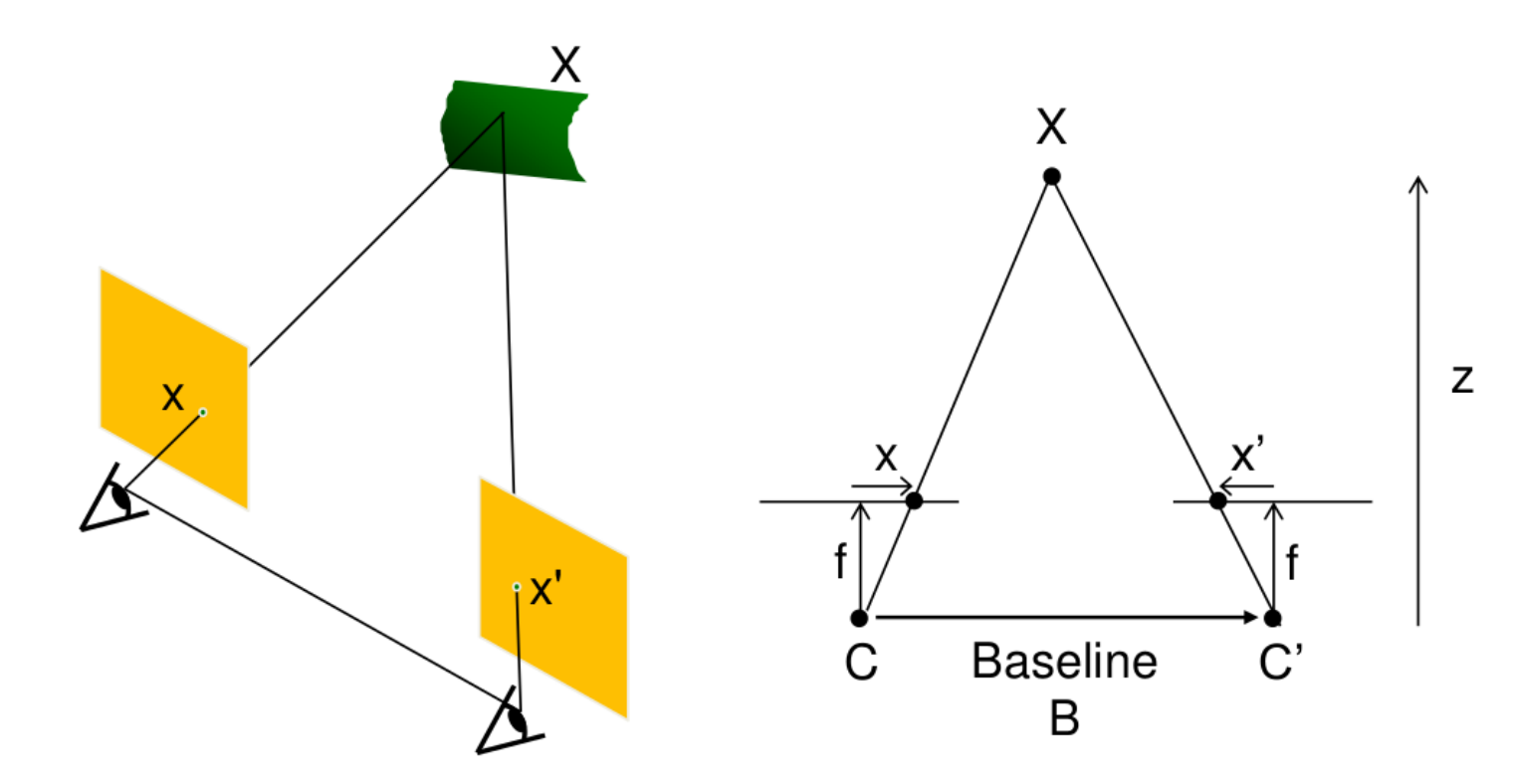
Søren Olsen

Department of Computer Science
University of Copenhagen

Topics for today's lecture

- Stereo vision, Epipolar geometry, Fundamental matrix etc
- Correspondence analysis
- Feature based vesus dense solutions
- Scale-space and Coarse-to-fine
- Reconstsructions from sparse data

Multiple View Correspondences

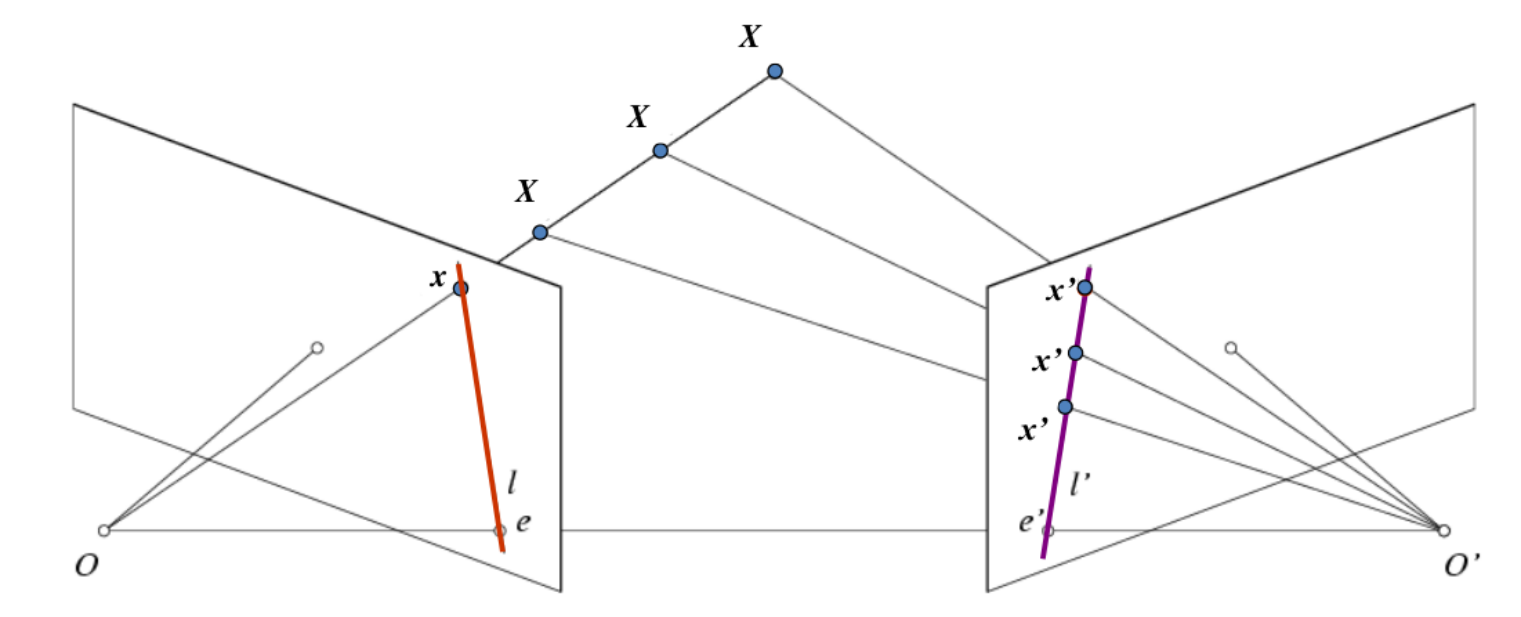


If we can recover x' from x we can recover depth: $z = -\frac{fB}{x'-x}$.

Image Correspondence

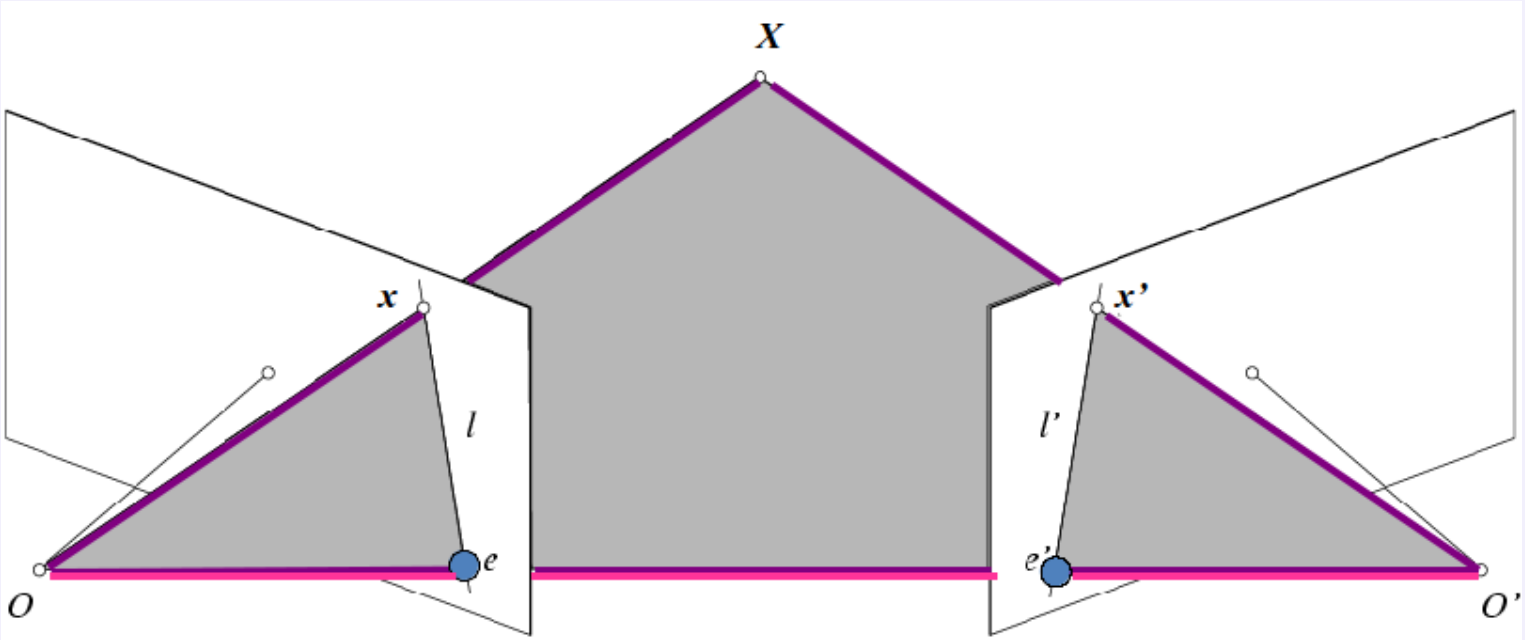
How do we match points from image 1 to image 2: You should have seen some of it before, but this is not the end of the story!

Epipolar Constraints



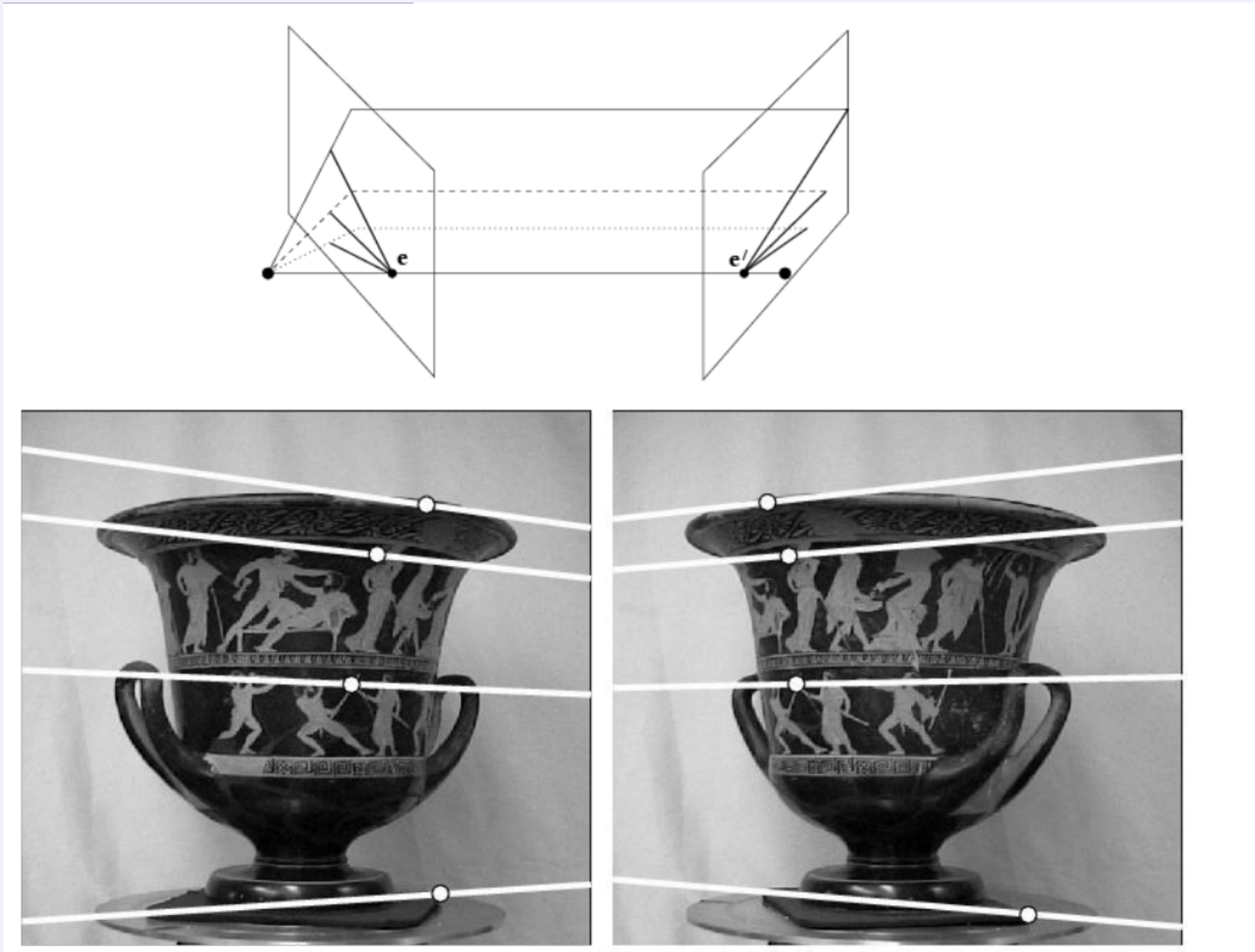
- Corresponding point for x must lie in corresponding line l'
- Corresponding point for x' must lie in corresponding line l

Epipolar Constraints

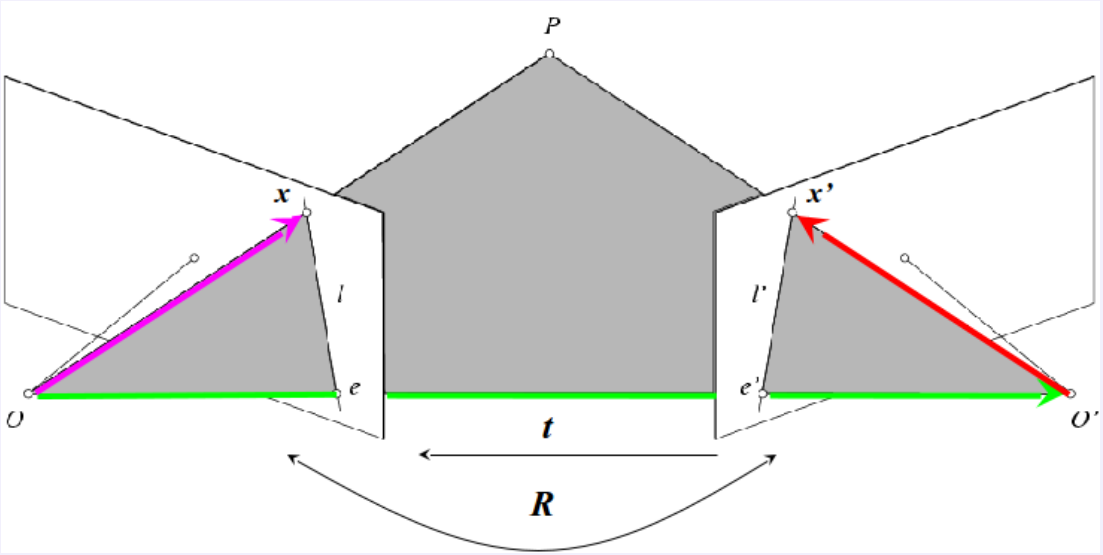


- Line connecting O and O': baseline
- Plane through baseline and x or x': Epipolar Plane
- Epipoles: intersection of baseline and image planes: projection of the other camera center.
- Epipolar Lines - intersections of epipolar plane with image planes (always come in corresponding pairs)

Example: Converging cameras



Calibrated Case



Camera parameters known for the two cameras: calibration matrices K and K'

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 y and y' are related by a 3×3 matrix E built from R and t

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

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

The fundamental matrix F

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

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

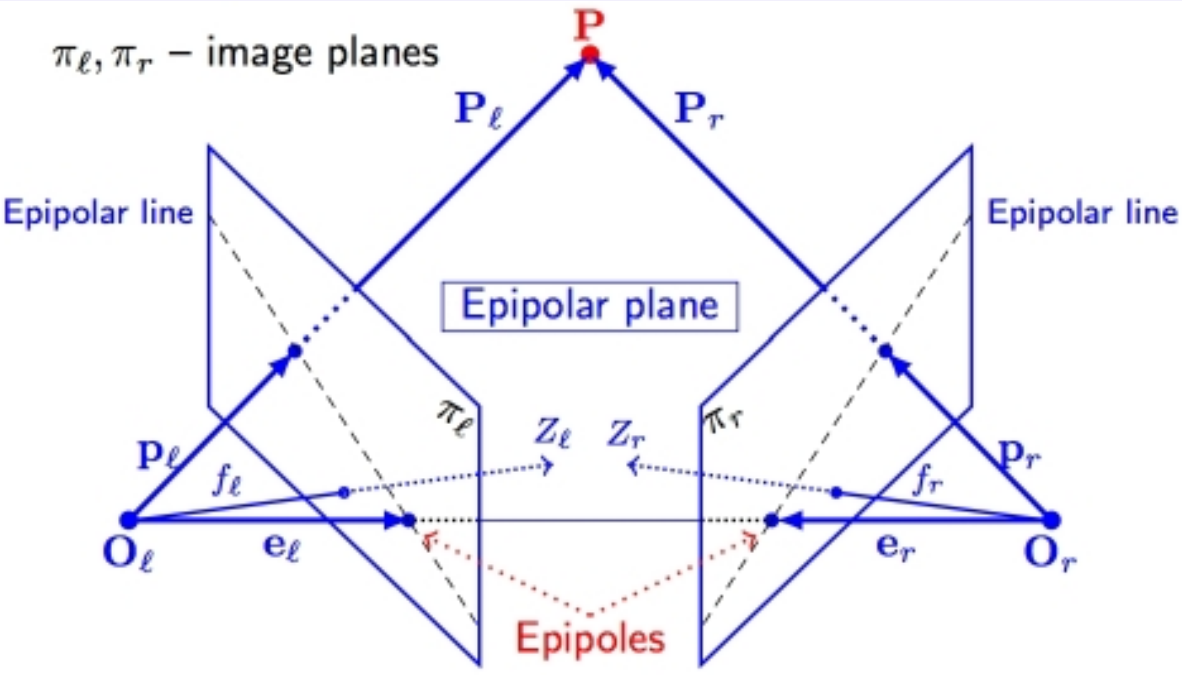
  where x and x' are the homogene representation of the corresponding points in image coordinates, and where F is the fundamental matrix.

- Given a sufficent 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.

Calibration and reconstruction

- Given (sufficient) image point correspondences, the fundamental matrix F may be estimated using linear algebra.
- F has only 7 degrees of freedom (not 9) because it is defined up to scaling only, and because $\det(F) = 0$.
- Linear estimation of F is easy, but not accurate. In practice a non-linear post-optimization is needed.
- Given F, the stereo correspondence problem is reduced to a one-dimensional search along the epipolar lines.
- Given projections (x, y) of known 3D points (X, Y, Z), the calibration matrix K may be estimates using linear algebra.
- Given F, K and K', reconstructions of 3D points is possible (using linear algebra) from image point correspondences.

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

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

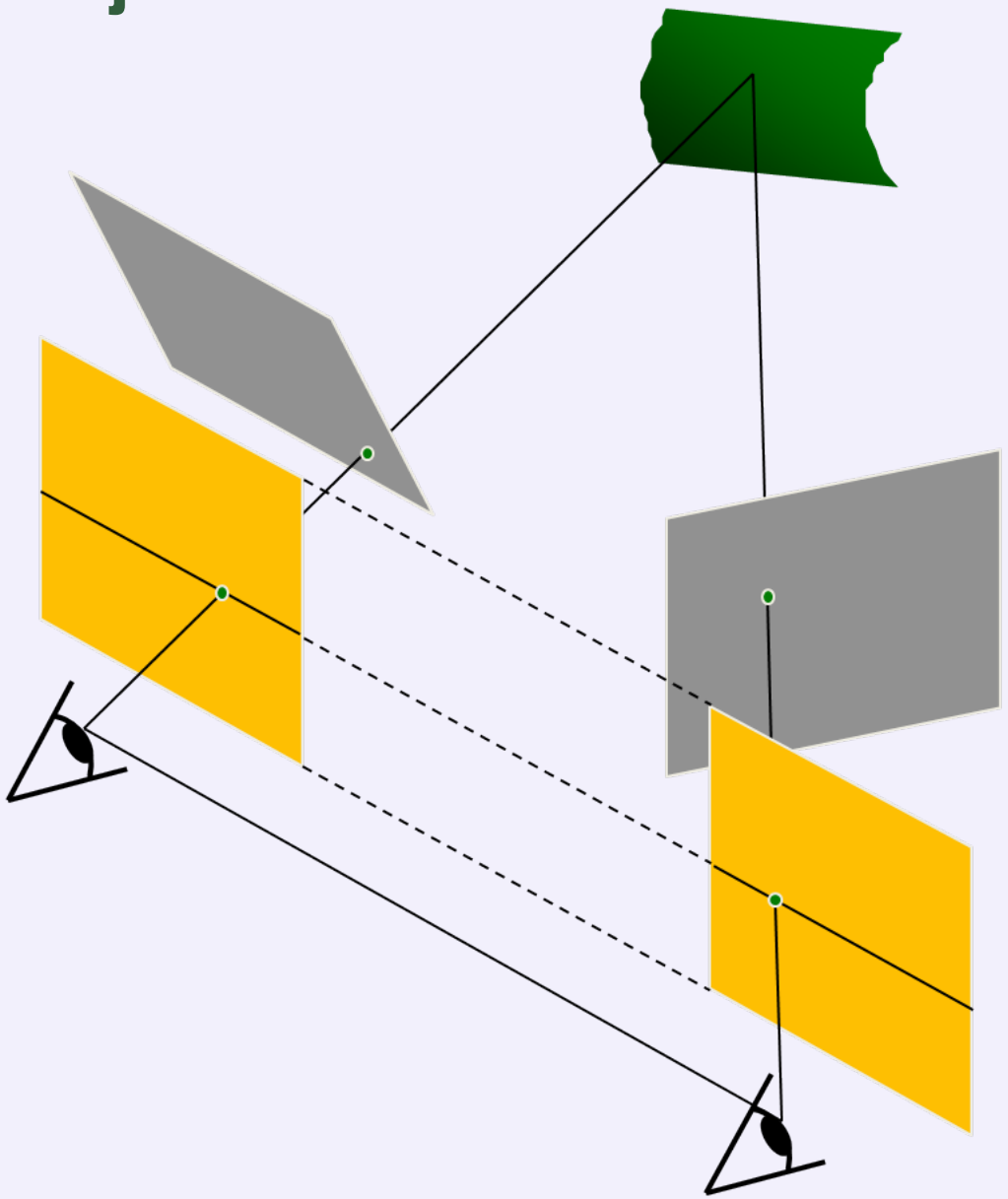
we finally get:

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

Non horizontal Scan lines

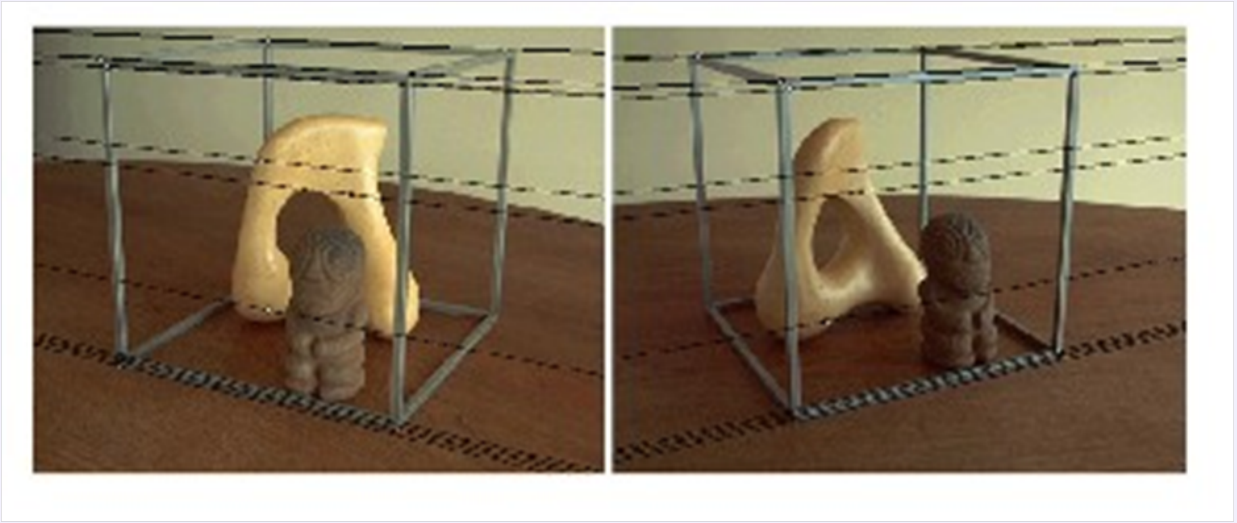
- If calibration known, the essential matrix provides epipolar constraints.
- What when cameras are in general position and calibration is unknown?
- Non calibrated views: Estimate the fundamental matrix.
- Knowing Essential or Fundamental matrix allows (almost) for image rectification.

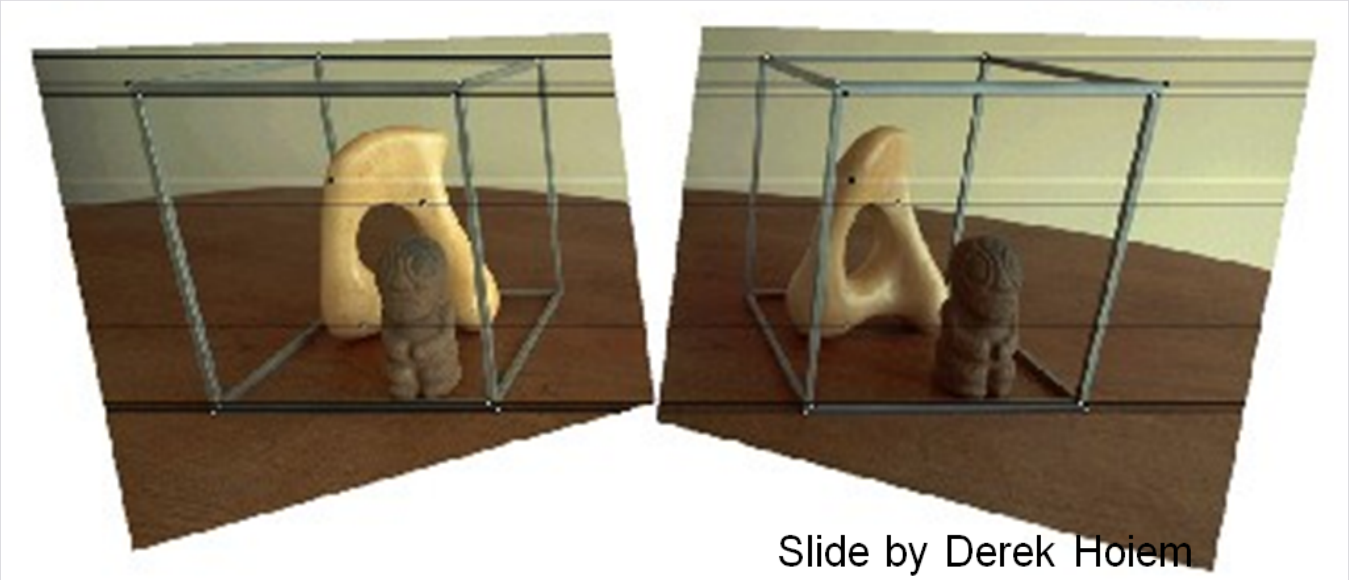
Projective Rectification



- 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





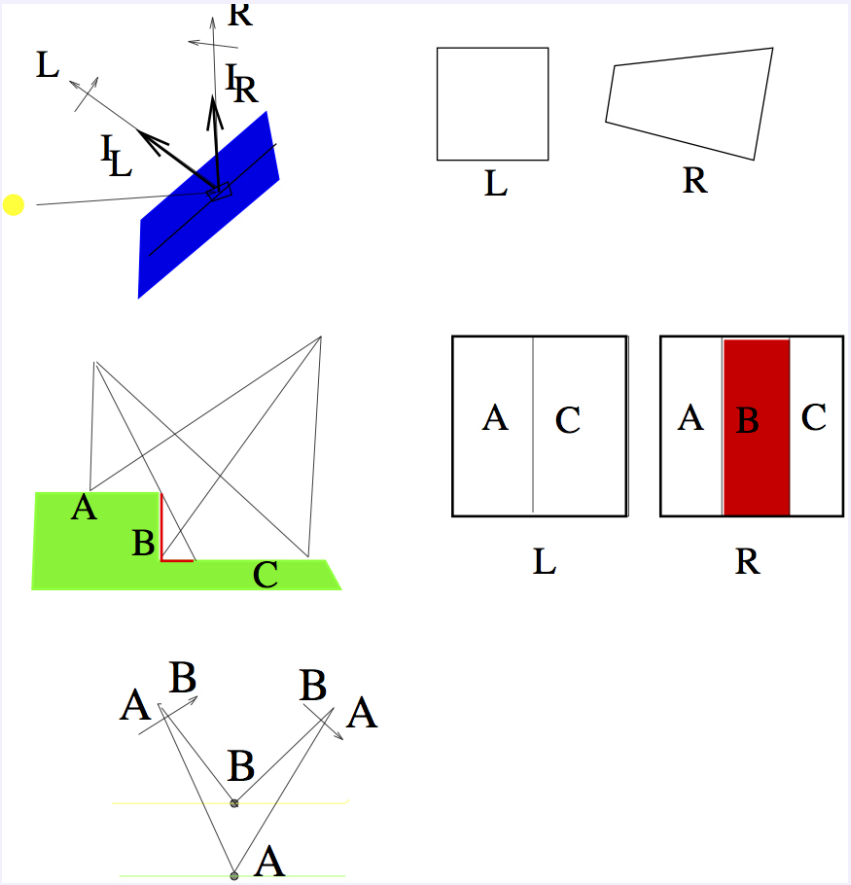
Correspondence analysis

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 alingnment/warping etc.
- Most mammalians have stereo vision
- Except for auto-focus cameras, stereo is the most widely applied passive technique for 3D-measurement.

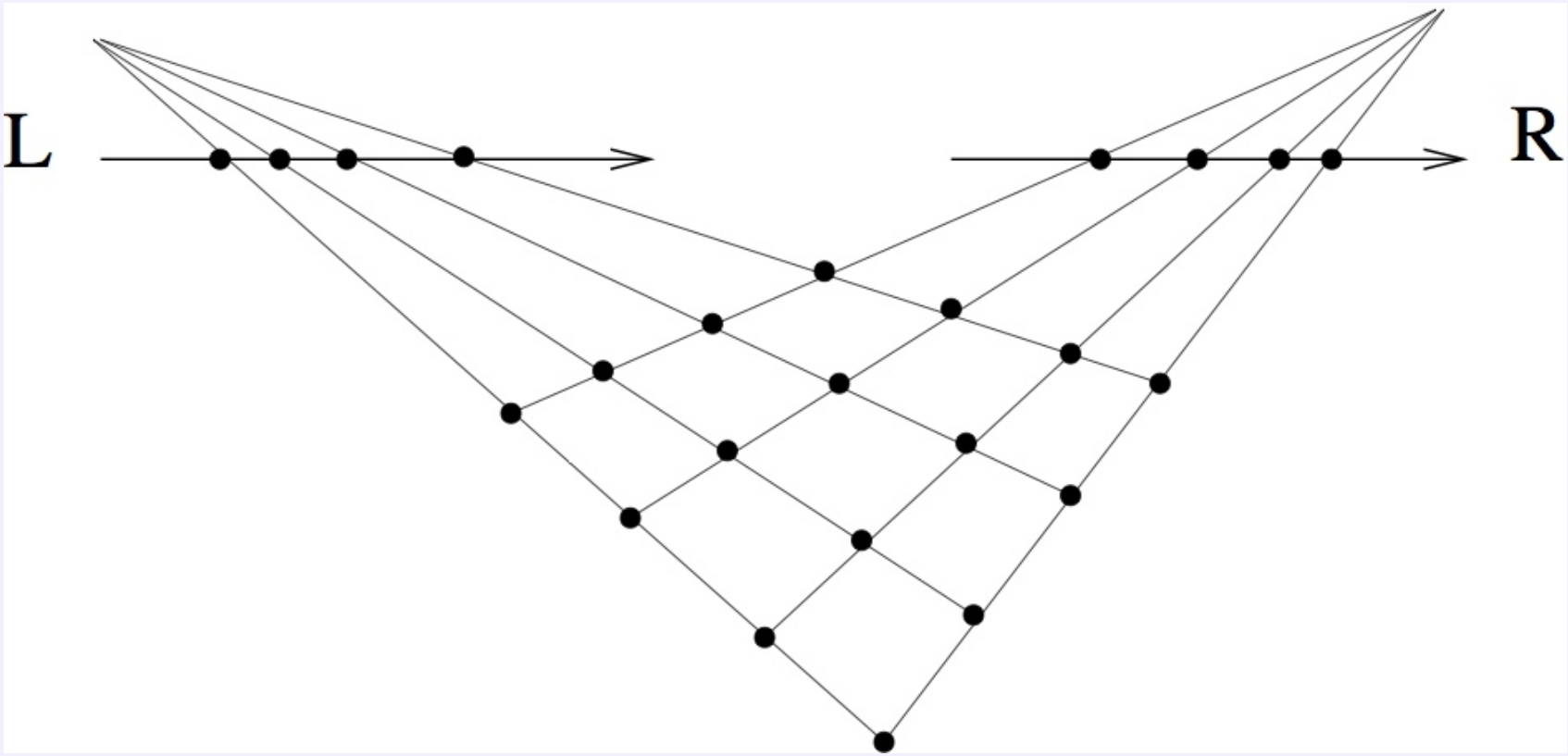
Problems

- 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 mathes exist in the other image ?



Simplifying assumptions

- Intensities are similar, eg. $|E_L - E_R| \leq \theta$ or are spartially 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

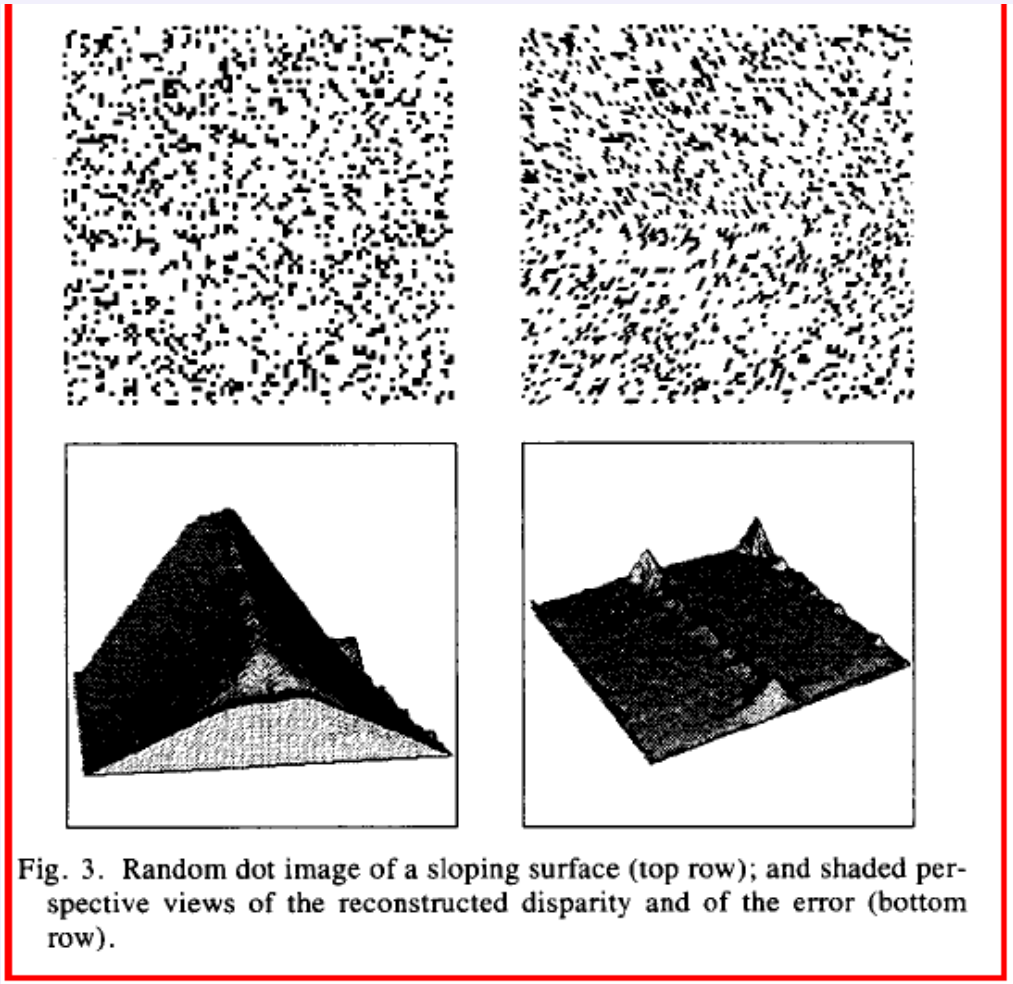


Fig. 3. Random dot image of a sloping surface (top row); and shaded perspective views of the reconstructed disparity and of the error (bottom row).

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

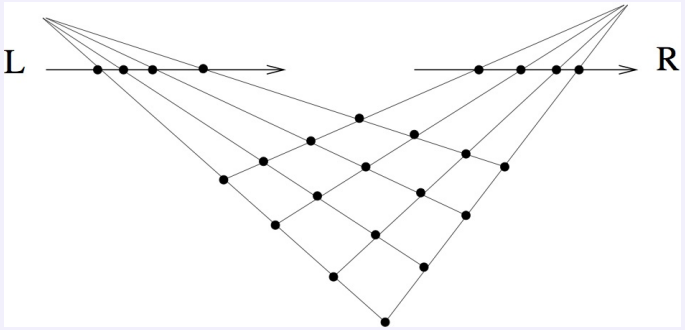
Example: Large disparity gradient 2

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 ?

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:
$N(n) = n!$. $N(4) = 12$.

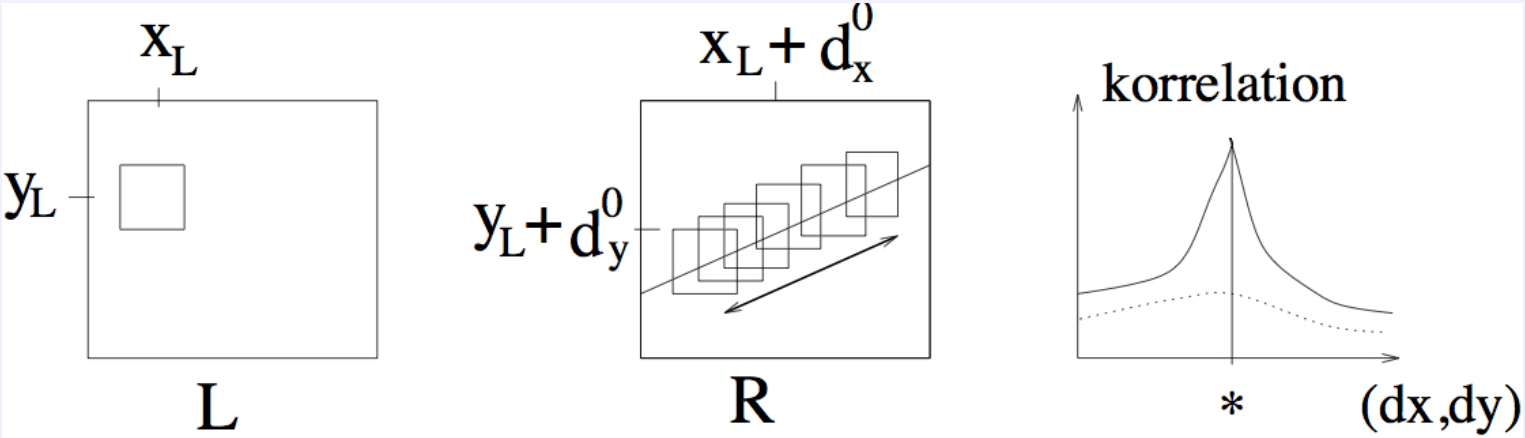
Assume strong ordering and uniqueness:
$N(n) = 1$.

Correspondence analysis

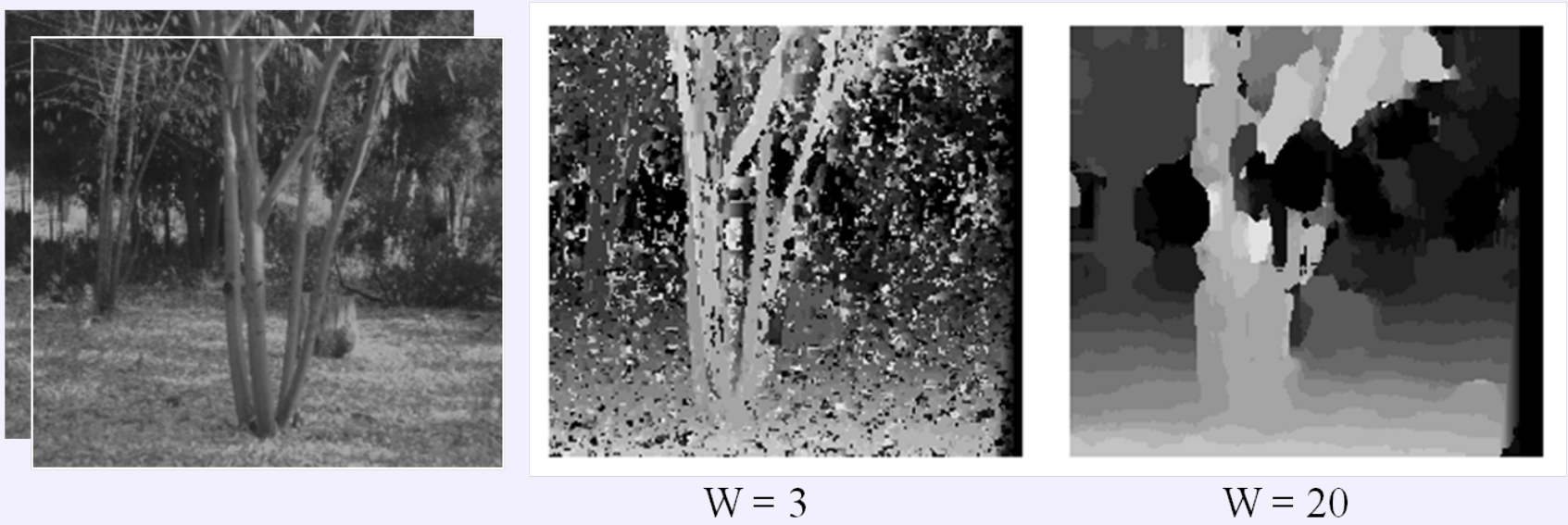
- Dense intensity based methods may be accurate but is very noise sensitive and have a small capture area. Also, they may be computationally expensive.
- Sparse feature based methods is faster, more reliable, and have larger capture area, but results in scattered depth information.
- Very local dense features as edge often has a short descriptor.
- 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.

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: Normalized 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 normalized 2D cross correlation is deined 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 thge 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 $(f \circ g)(x, y)$.

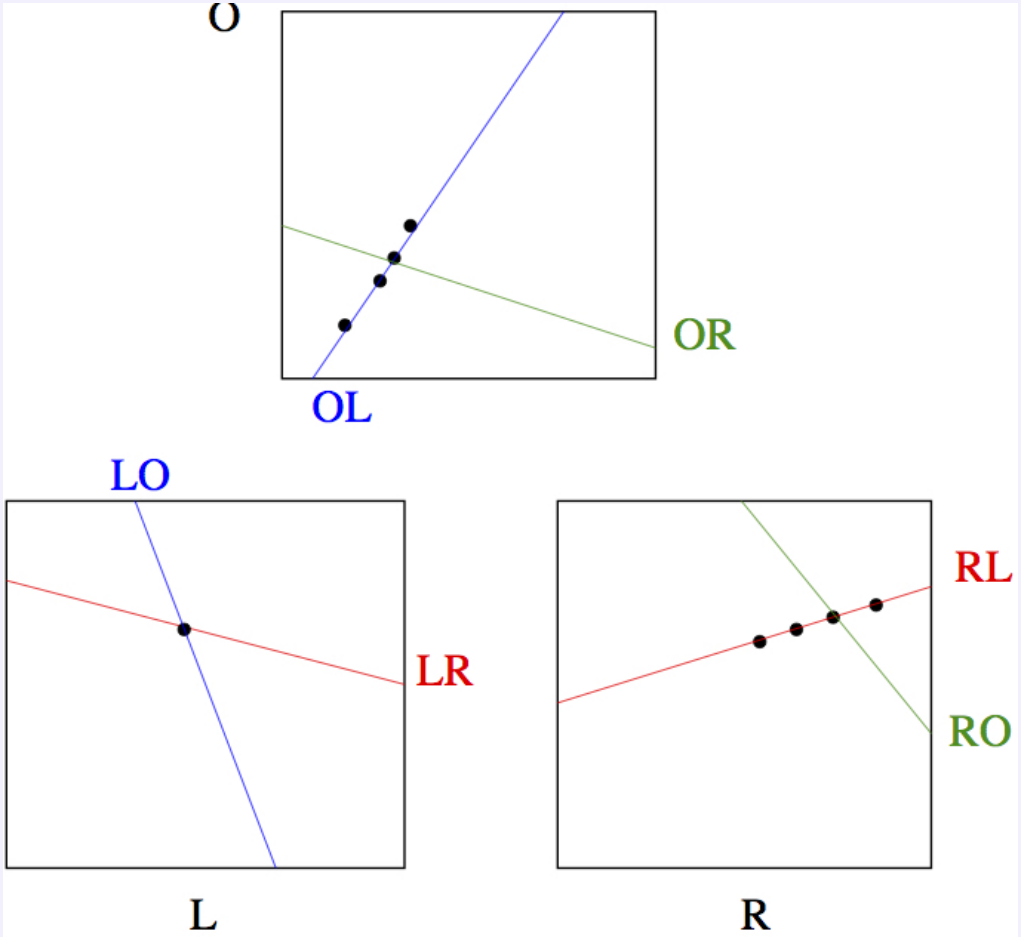
Feature base stereo

- Which feature ?
- For depth reconstructions, the density of points should be maximized. Often edge points localized with sub-pixel accuracy is chosen.
- For matching efficiency, the quality (disambiguation power) of the descriptor should be maximized. This is computational more expensive. Often corners or blobs with SIFT/HOG-like descriptors are used.

In practice hybrid systems may be used: Sparse descriptive feature points are used for fundamental matrix estimation. Edge points are used to obtain an initial reconstruction. Finally, intensity based methods are used for fine-tuning and filling-in.

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 analyze images. In practice we represent each image at a number of scales.

Please check previous slides for the definition and details on scale-space.

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

Coarse-to-fine

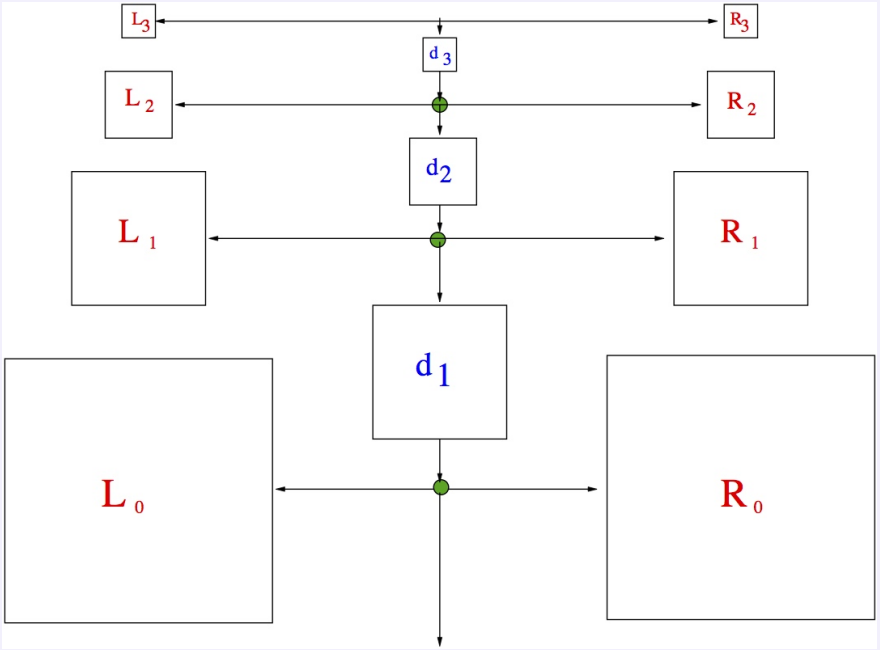
Pyramid-based coarse to fine approaches:

- 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

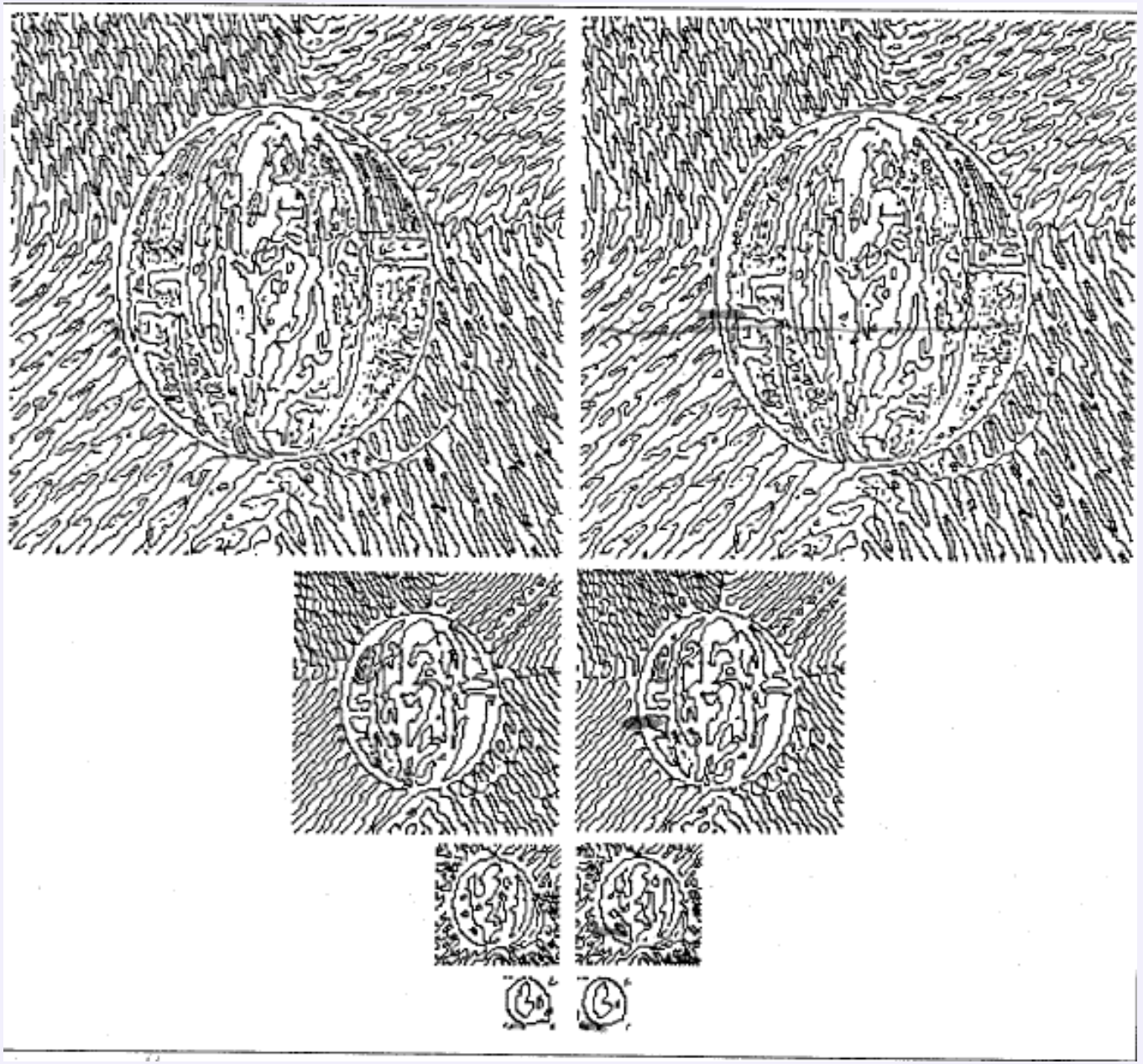
In practice the disparity may be large (several hundred pixels) and have large variation (eg. from -50 to +50 pixels). To reduce the size of the search area we need an estimate of the disparity.



Method: Succesive smoothing and downsampling

The total space requirement is:

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



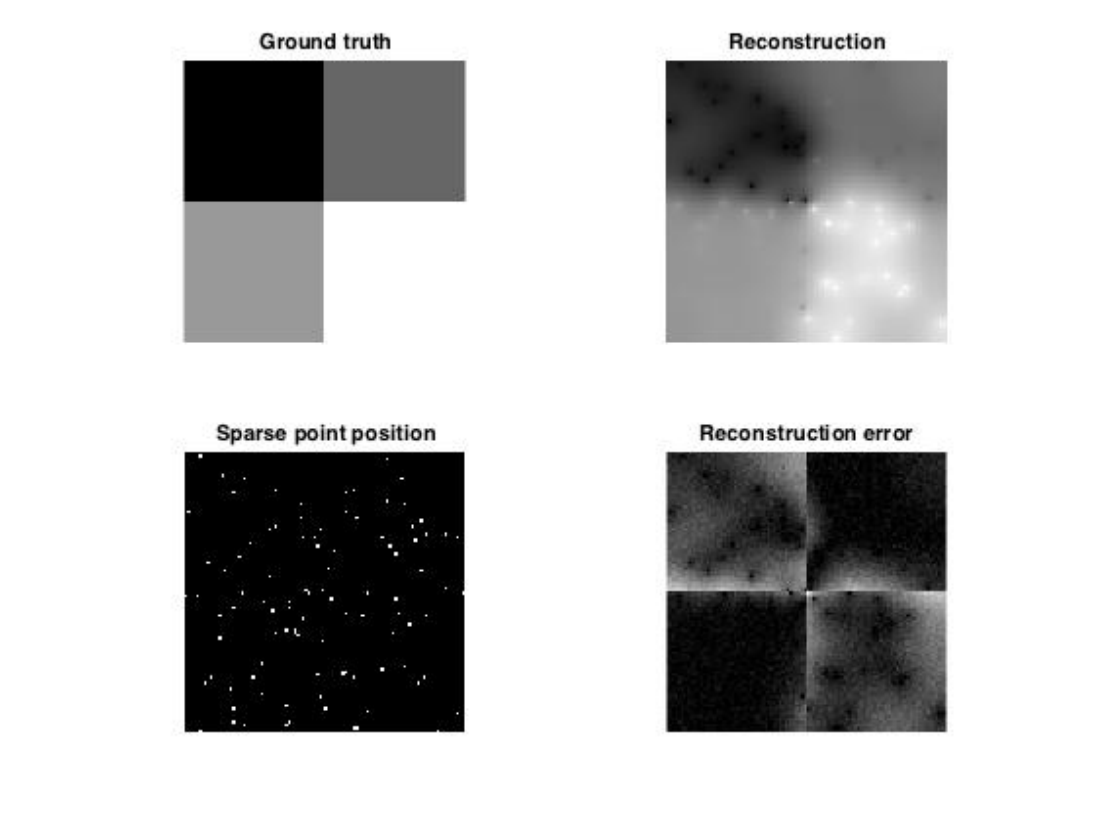
Surface interpolation

It is possible to fit a thin plate spline surface to the estimated sparse depth measurements. Usually, an (slow) iterative updating is used. To speed up a simple and fast method is used to obtain a initial estimate.

Surface interpolation is too advanced for this course.

Reconstruction on few data

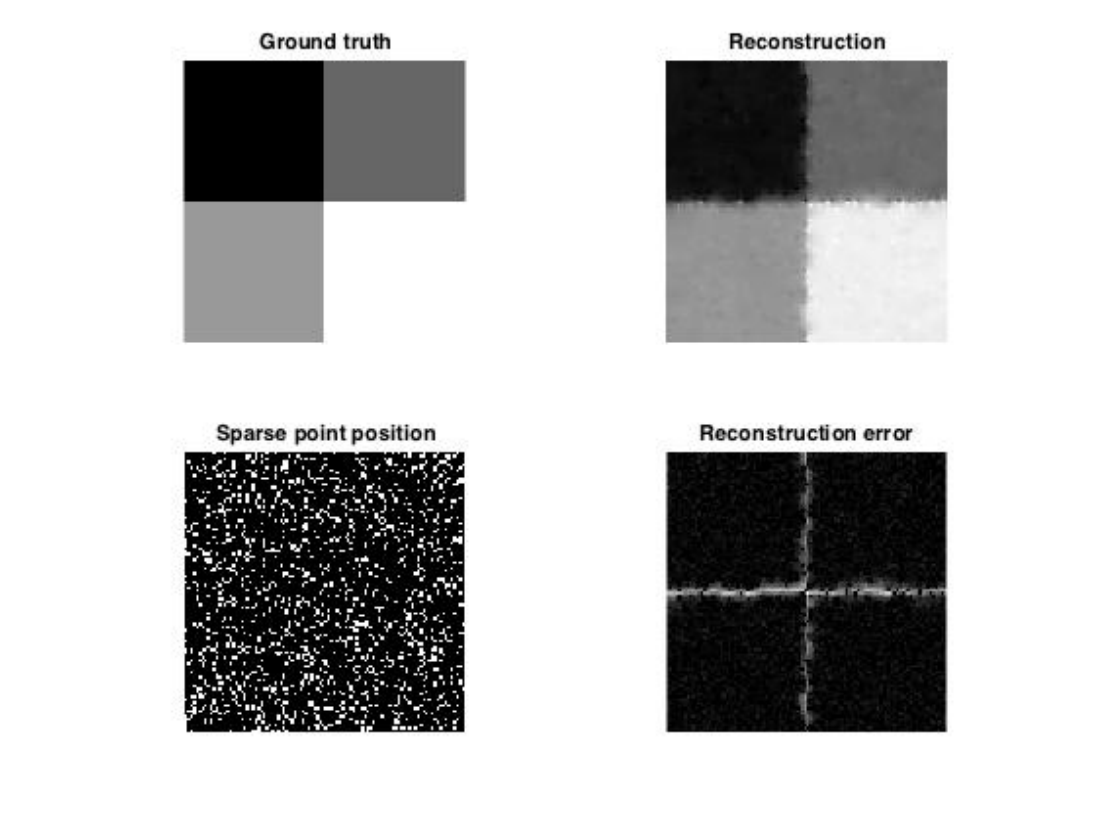
The results below are produced by the MATLAB-program interp(.) that you may use in assignment 3.



Very sparse data, like SIFT-points makes reconstruction hard.

Reconstruction on more data

Dense data like edge points makes reconstruction better.



Discontinuities and occlusion

- It is possible to perform discontinuous regularization, where the smoothmess term is disregarded when the surface is bended more than some threshold.
- Discontinuities are accompanied by occluded areas. If any point in an occluded area is matched it will be wrong. If not, the surface will be reconstructed more smooth than what it should.

Example

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.

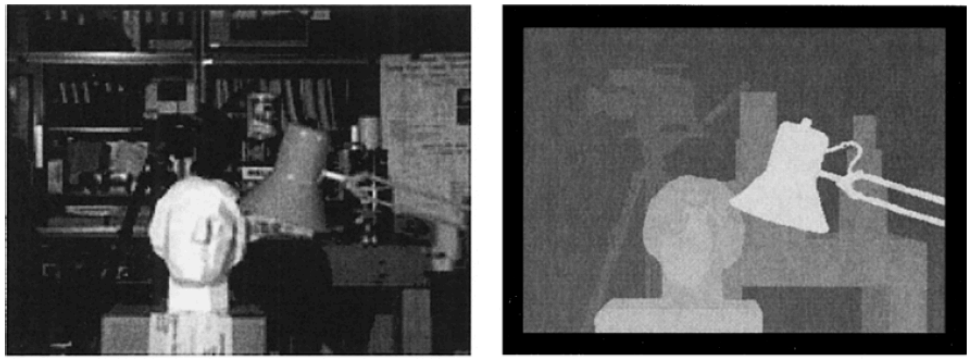


Figure 4. Segmented region maps: (a) original image, (b) true disparities, (c) textureless regions (white) and occluded regions (black), (d) depth discontinuity regions (white) and occluded regions (black).

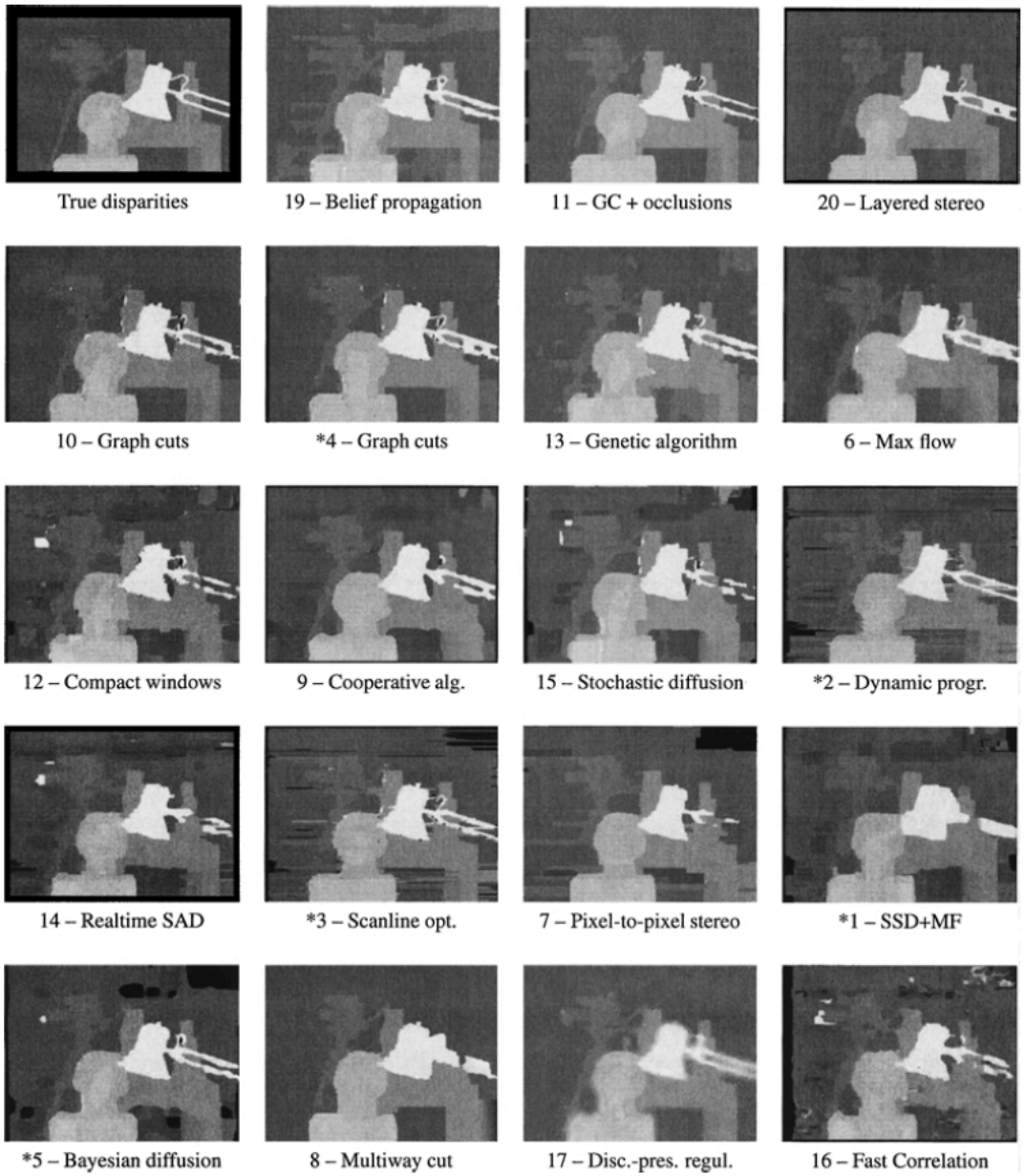


Figure 17. Comparative results on the Tsukuba images. The results are shown in decreasing order of overall performance ($B_{\bar{O}}$). Algorithms implemented by us are marked with a star.

Pentagon

Output of 110-lines of code programs to verify assignment: