Introduction to computer vision XI

Instructors: Jean Ponce and Matthew Trager
jean.ponce@inria.fr, matthew.trager@cims.nyu.edu

TAs: Jiachen (Jason) Zhu and Sahar Siddiqui
jiachen.zhu@nyu.edu, ss12414@nyu.edu

Slides will be available after class at:
https://mtrager.github.io/introCV-fall2019/

Stereo
- Essential and fundamental matrices
- 8-point agorithm
- Rectification
- Triangulation
- Fusion algorithms

Structure from motion
- Problem definition
- Ambiguities
- Euclidean SFM from the essential matrix
- Affine SFM from two views
- Affine SFM from multiple views
- Projective SFM

Problem with eight-point algorithm

250906.36	183269.57	921.81	200931.10	146766.13	738.21	272.19	198.81
2692.28	131633.03	176.27	6196.73	302975.59	405.71	15.27	746.79
416374.23	871684.30	935.47	408110.89	854384.92	916.90	445.10	931.81
191183.60	171759.40	410.27	416435.62	374125.90	893.65	465.99	418.65
48988.86	30401.76	57.89	298604.57	185309.58	352.87	846.22	525.15
164786.04	546559.67	813.17	1998.37	6628.15	9.86	202.65	672.14
116407.01	2727.75	138.89	169941.27	3982.21	202.77	838.12	19.64
135384.58	75411.13	198.72	411350.03	229127.78	603.79	681.28	379.48

$$\begin{pmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \end{pmatrix} = - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

- Poor numerical conditioning
- Can be fixed by rescaling the data

The Normalized Eight-Point Algorithm (Hartley, 1995)

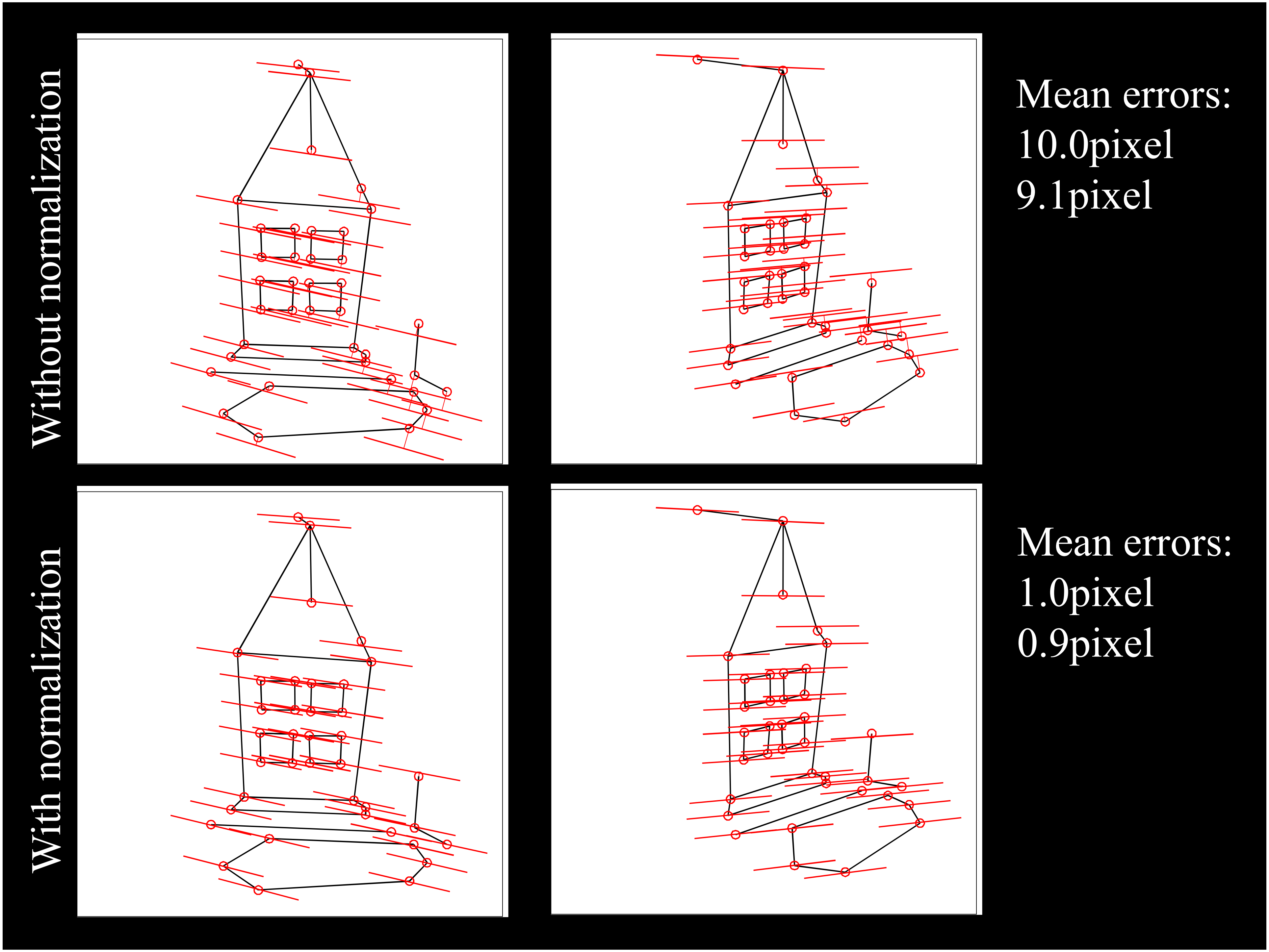
- Center the image data at the origin, and scale it so the mean squared distance between the origin and the data points is 2 pixels: $q_i = T p_i$, $q'_i = T' p'_i$
- Use the eight-point algorithm to compute F from the points q_i and q'_i.
- Enforce the rank-2 constraint.
- Output $T^T F T'$.

Non-Linear Least-Squares Approach (Luong et al., 1993)

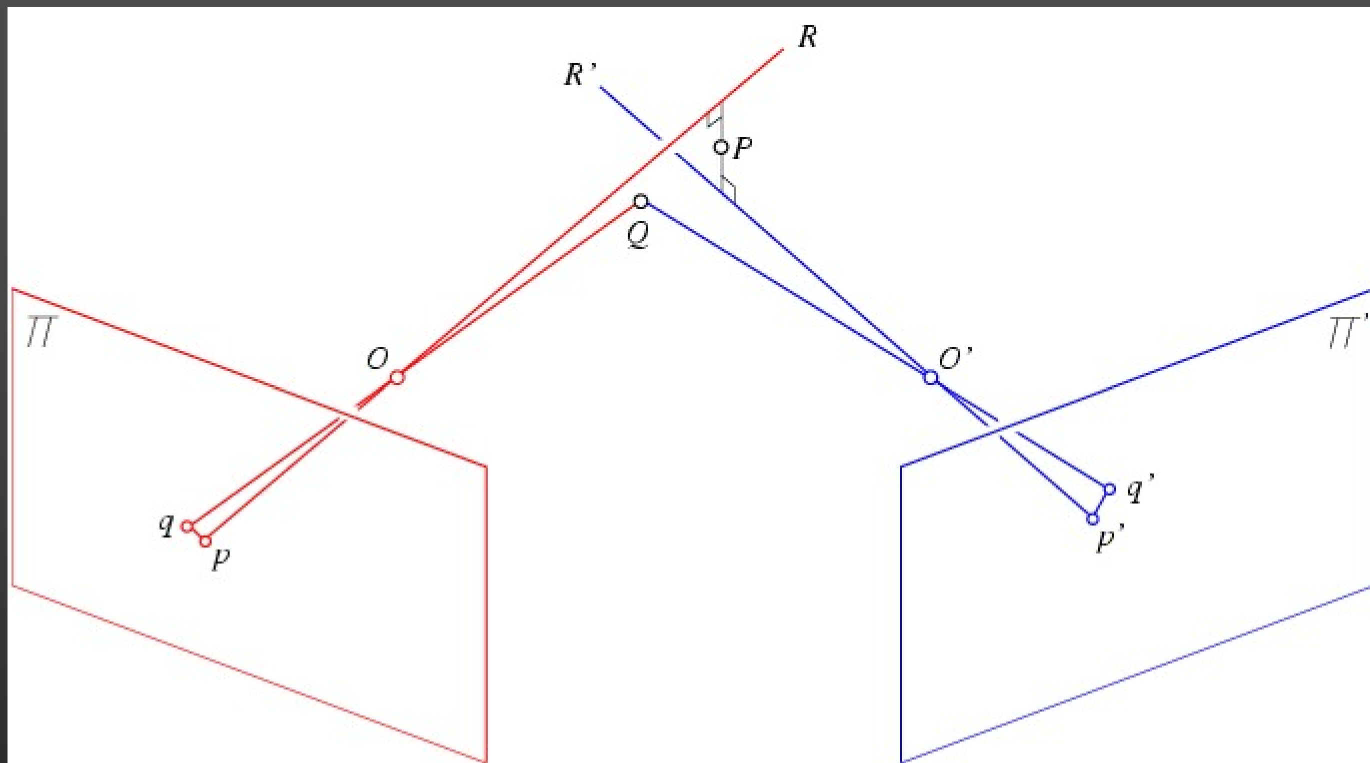
Minimize

$$\sum_{i=1}^{n} [\mathrm{d}^2(\boldsymbol{p}_i, \mathcal{F}\boldsymbol{p}'_i) + \mathrm{d}^2(\boldsymbol{p}'_i, \mathcal{F}^T \boldsymbol{p}_i)]$$

with respect to the coefficients of F , using an appropriate rank-2 parameterization.


Without normalization
Mean errors:
10.0pixel
9.1pixel
With normalization
Mean errors:
1.0pixel
0.9pixel

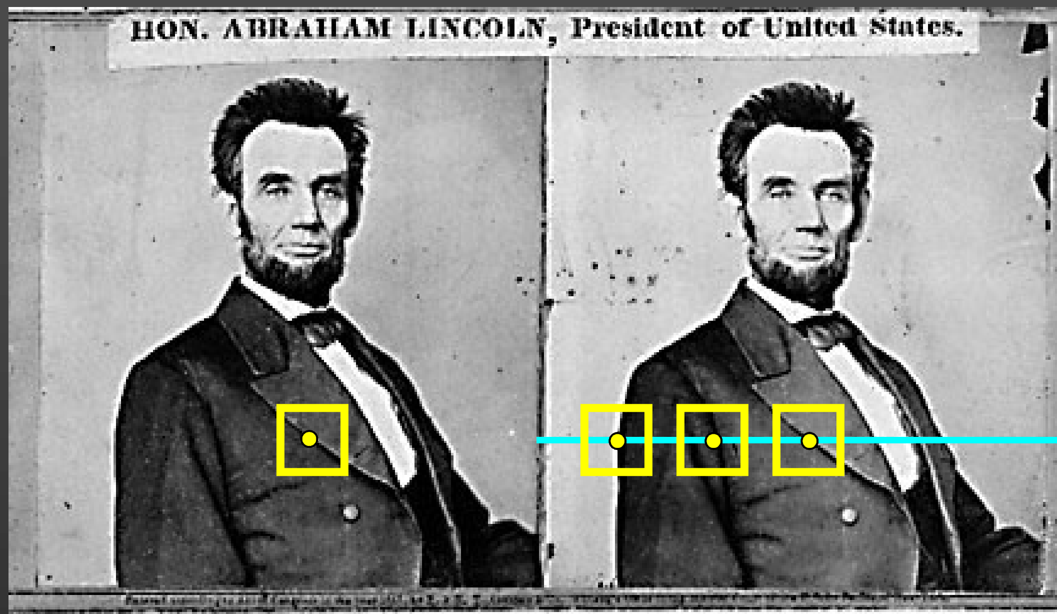
Reconstruction



- Linear Method: find P such that $\begin{cases} \boldsymbol{p} \times \mathcal{M}\boldsymbol{P} = 0 \\ \boldsymbol{p}' \times \mathcal{M}'\boldsymbol{P} = 0 \end{cases} \Longleftrightarrow \begin{pmatrix} [\boldsymbol{p}_\times]\mathcal{M} \\ [\boldsymbol{p}'_\times]\mathcal{M}' \end{pmatrix} \boldsymbol{P} = 0$

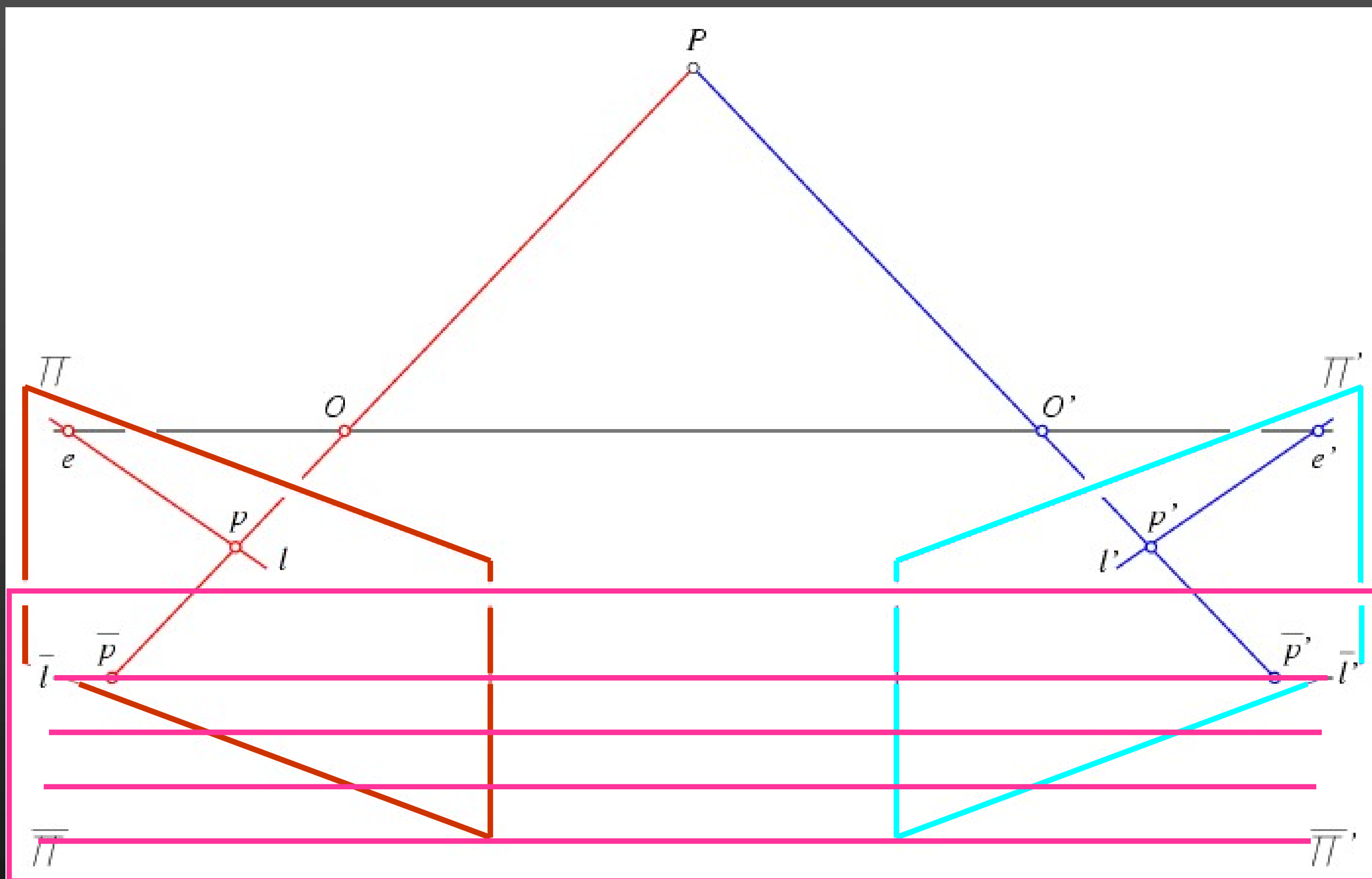
- Non-Linear Method: find Q minimizing $d^2(p,q) + d^2(p',q')$

Basic stereo matching algorithm



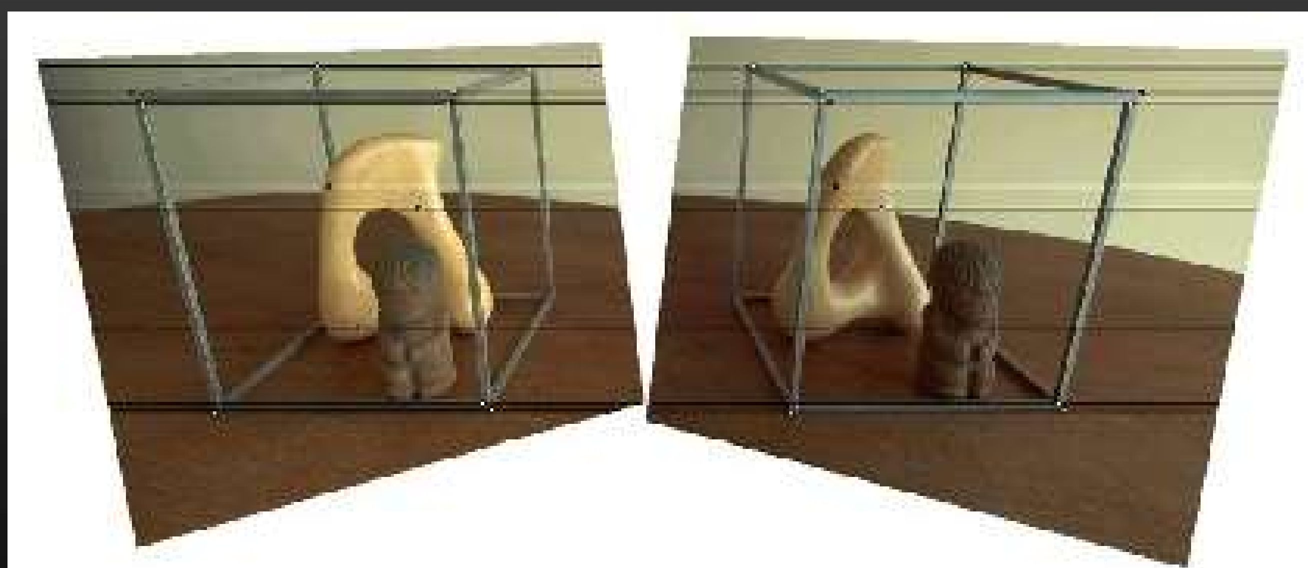
- For each pixel in the first image
  - Find corresponding epipolar line in the right image
  - Examine all pixels on the epipolar line and pick the best match
  - Triangulate the matches to get depth information
- Simplest case: epipolar lines are scanlines
  - When does this happen?

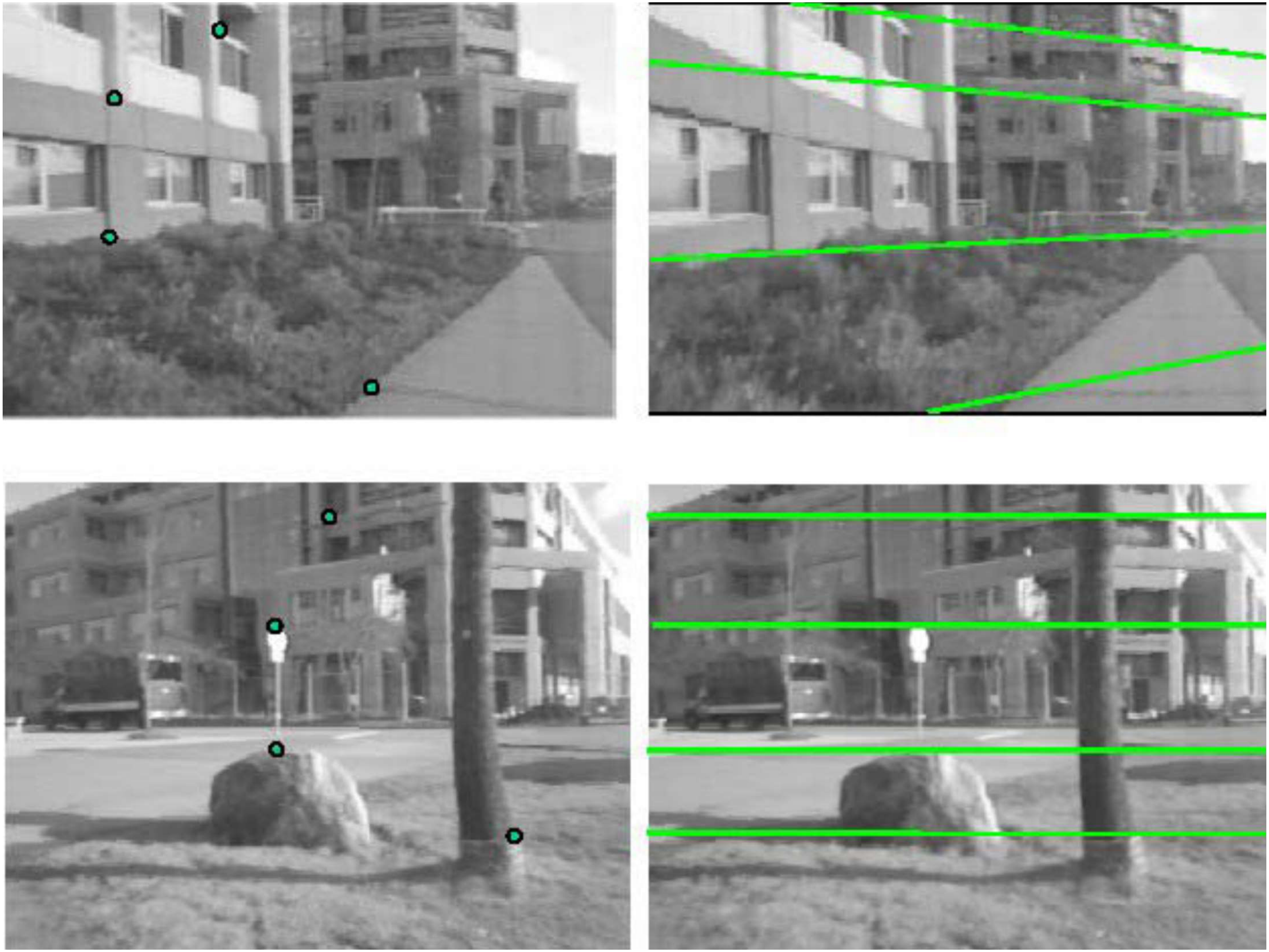
Rectification



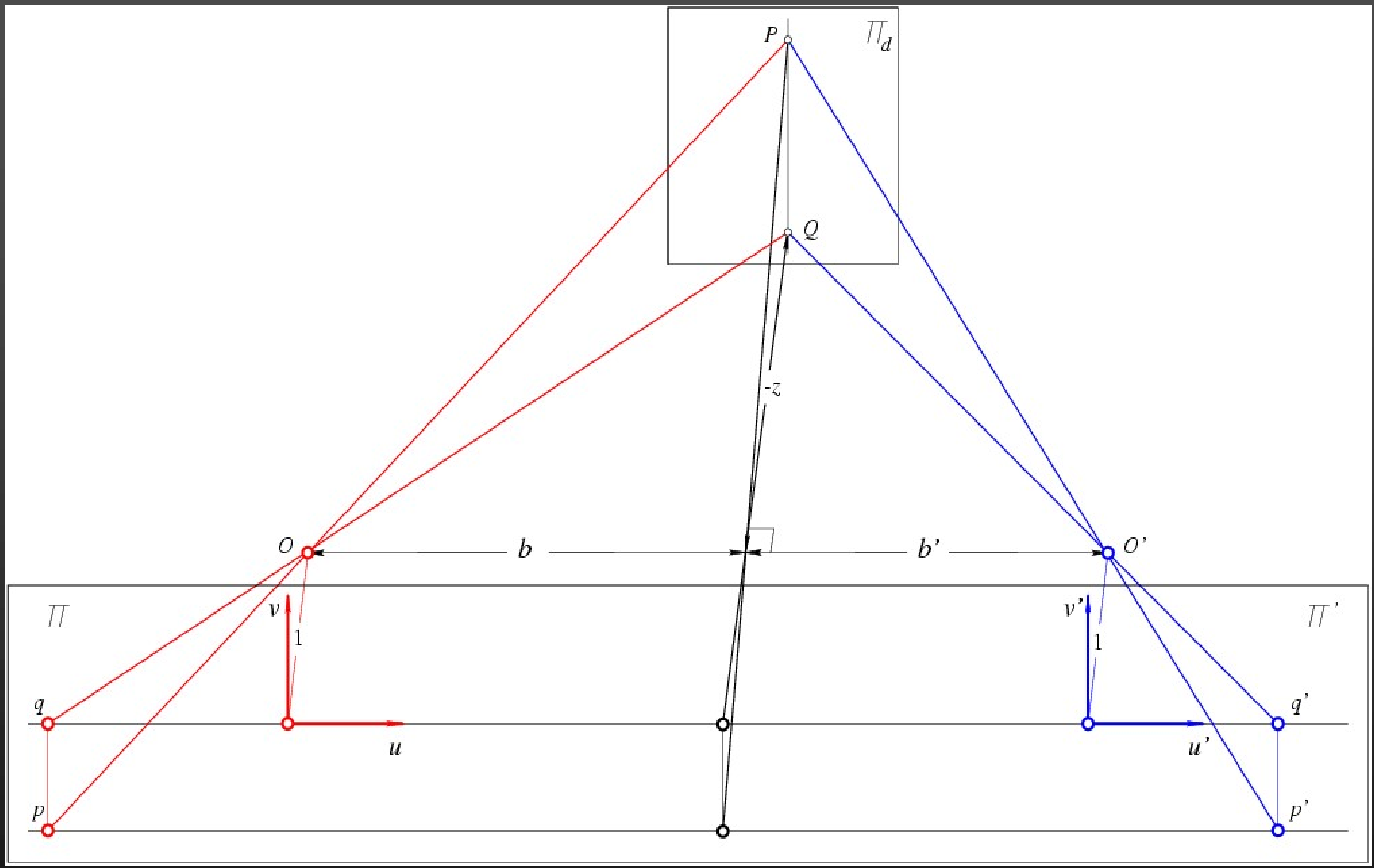
All epipolar lines are parallel in the rectified image plane.

Rectification example



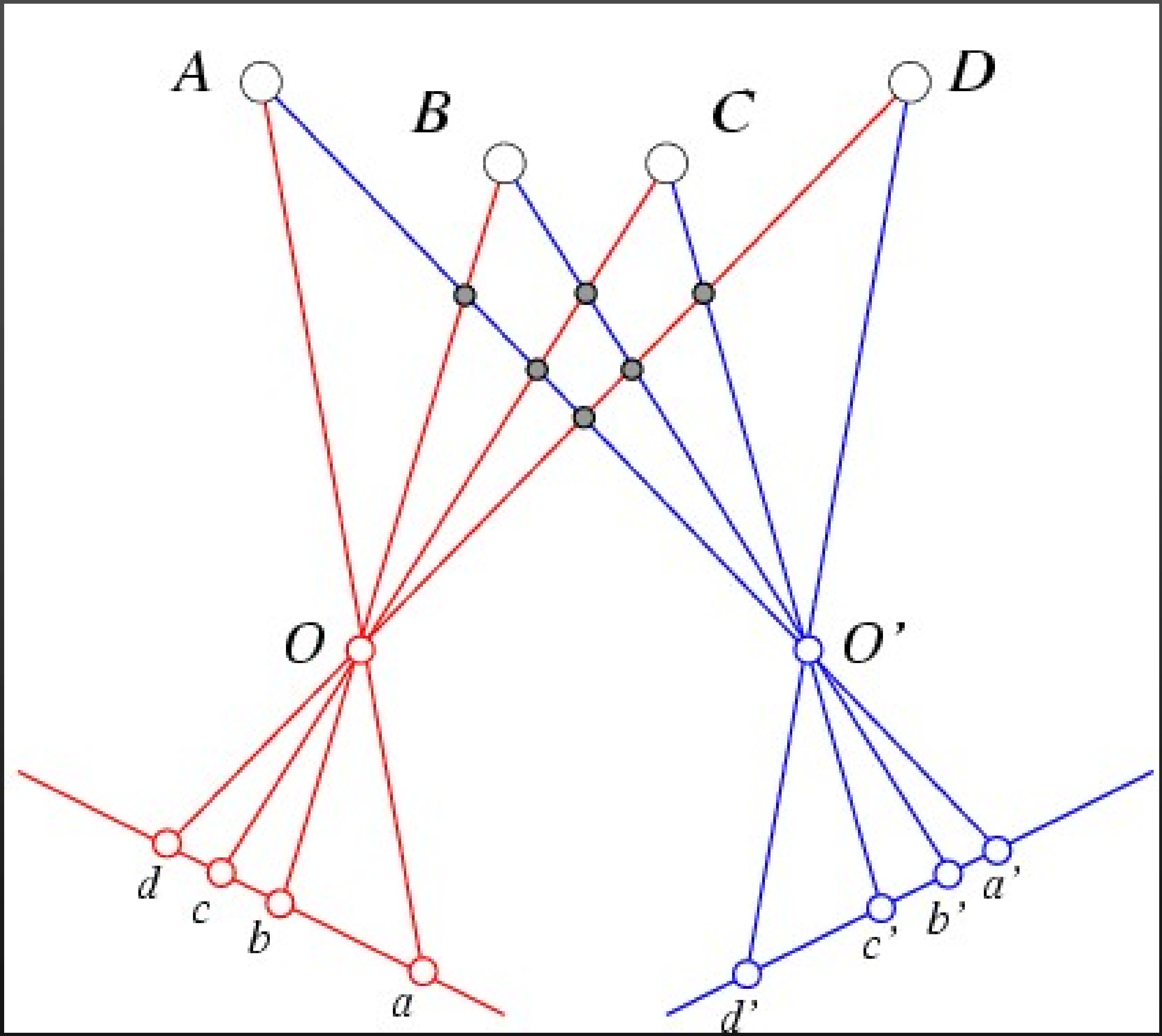


Reconstruction from Rectified Images

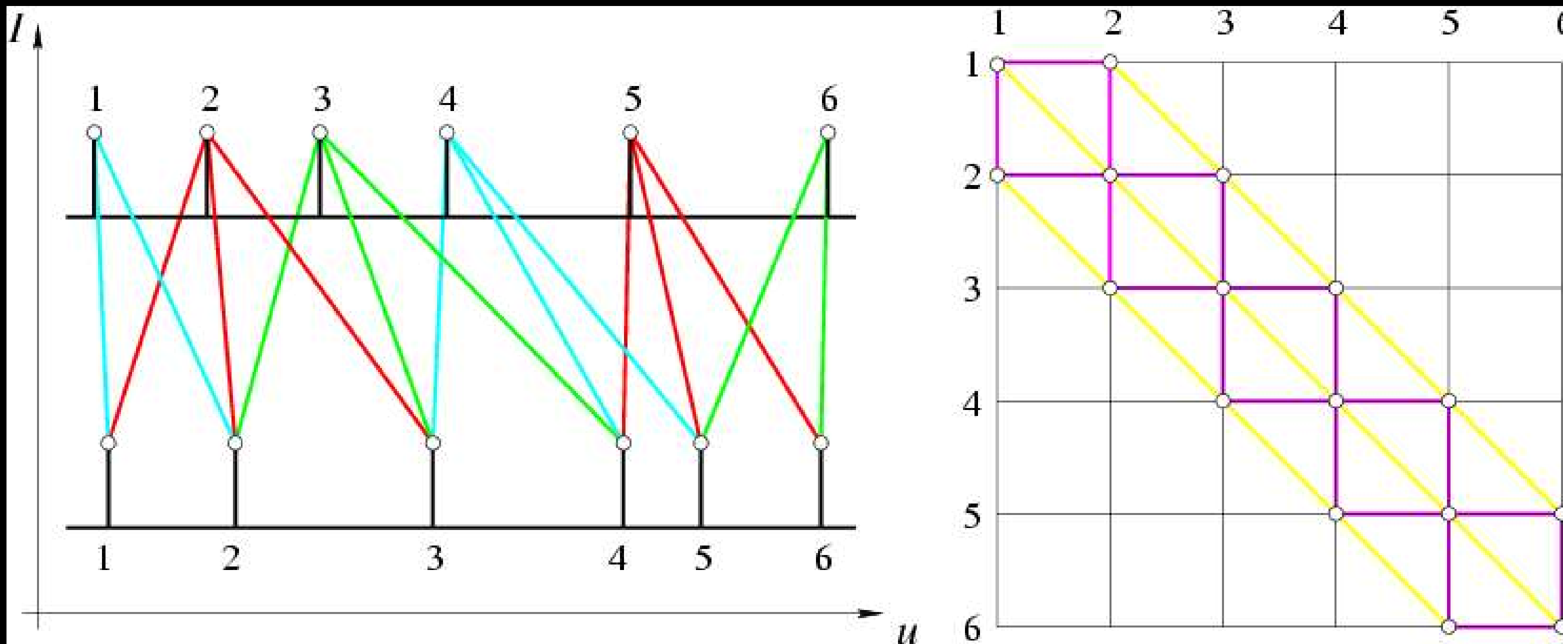


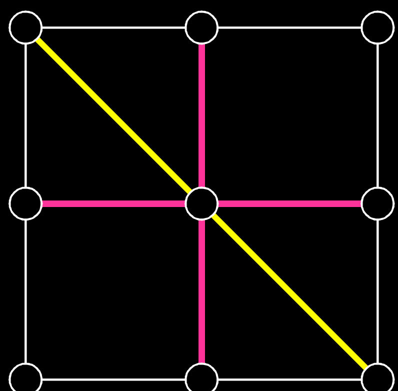
Disparity: d=u'-u. → Depth: z = -B/d.

Binocular fusion: a problem of correspondence



A Cooperative Model (Marr and Poggio, 1976)

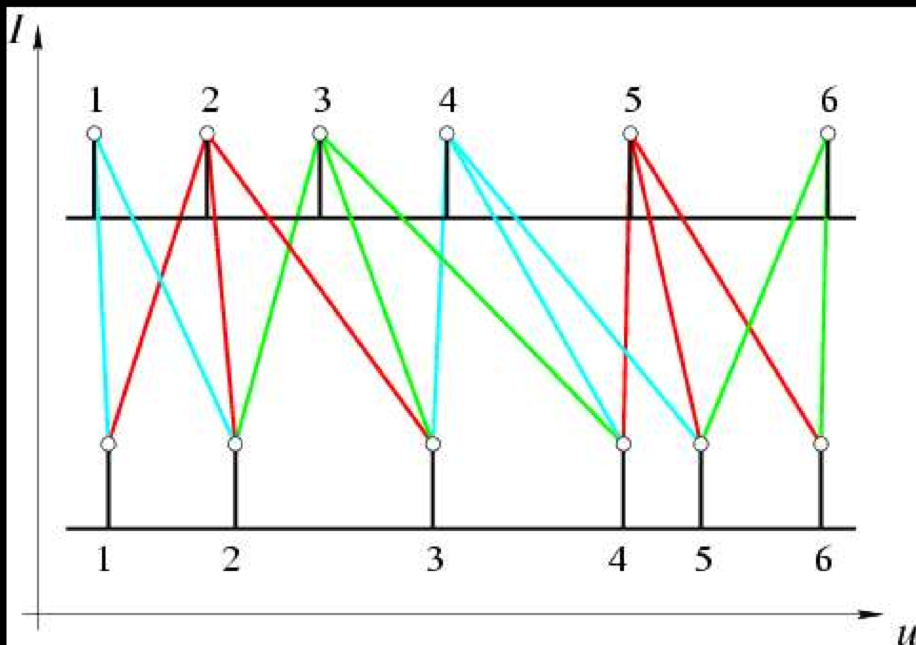


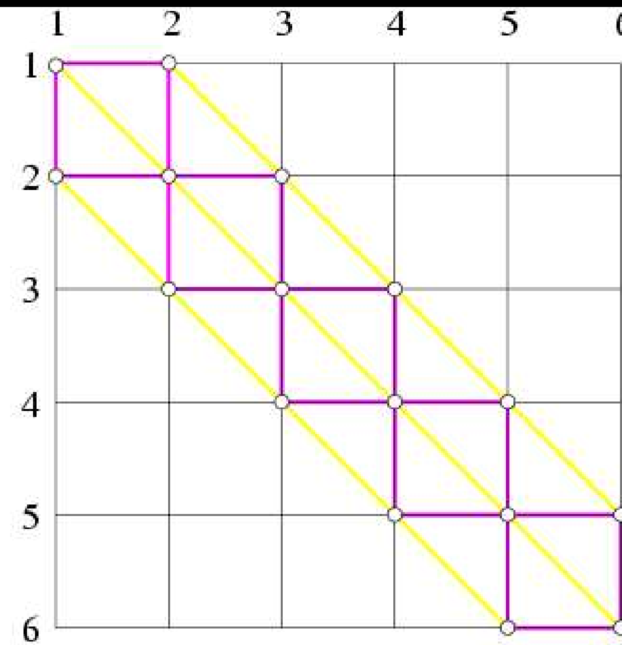


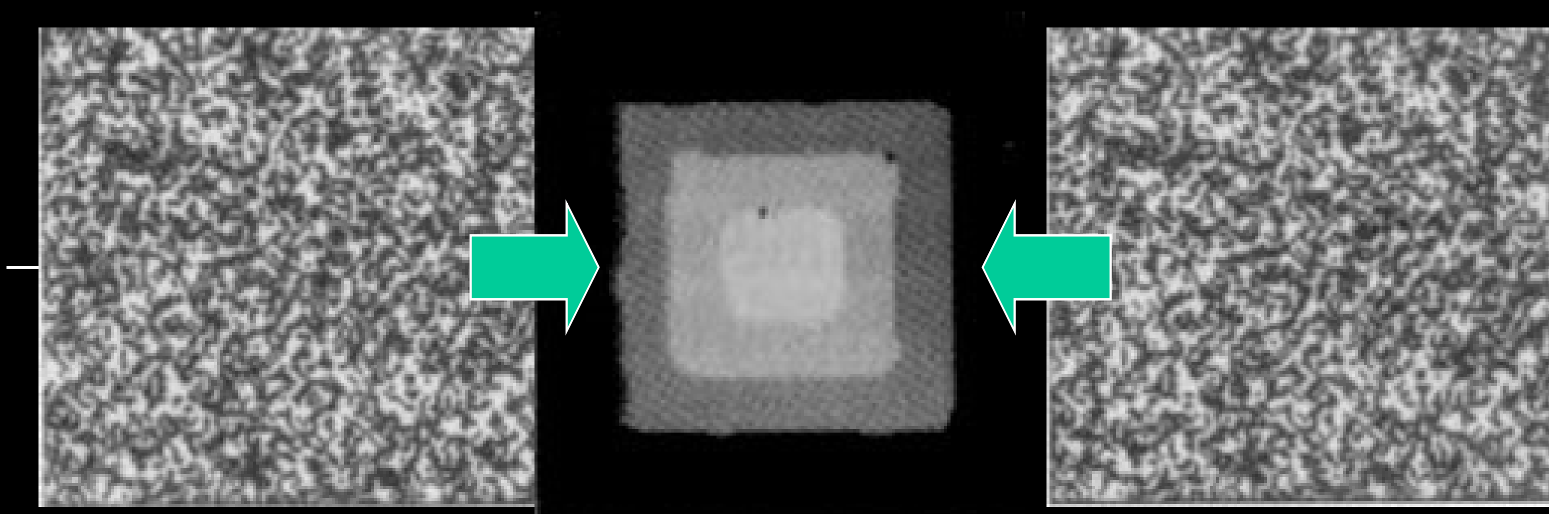
Excitory connections: continuity

Inhibitory connections: uniqueness

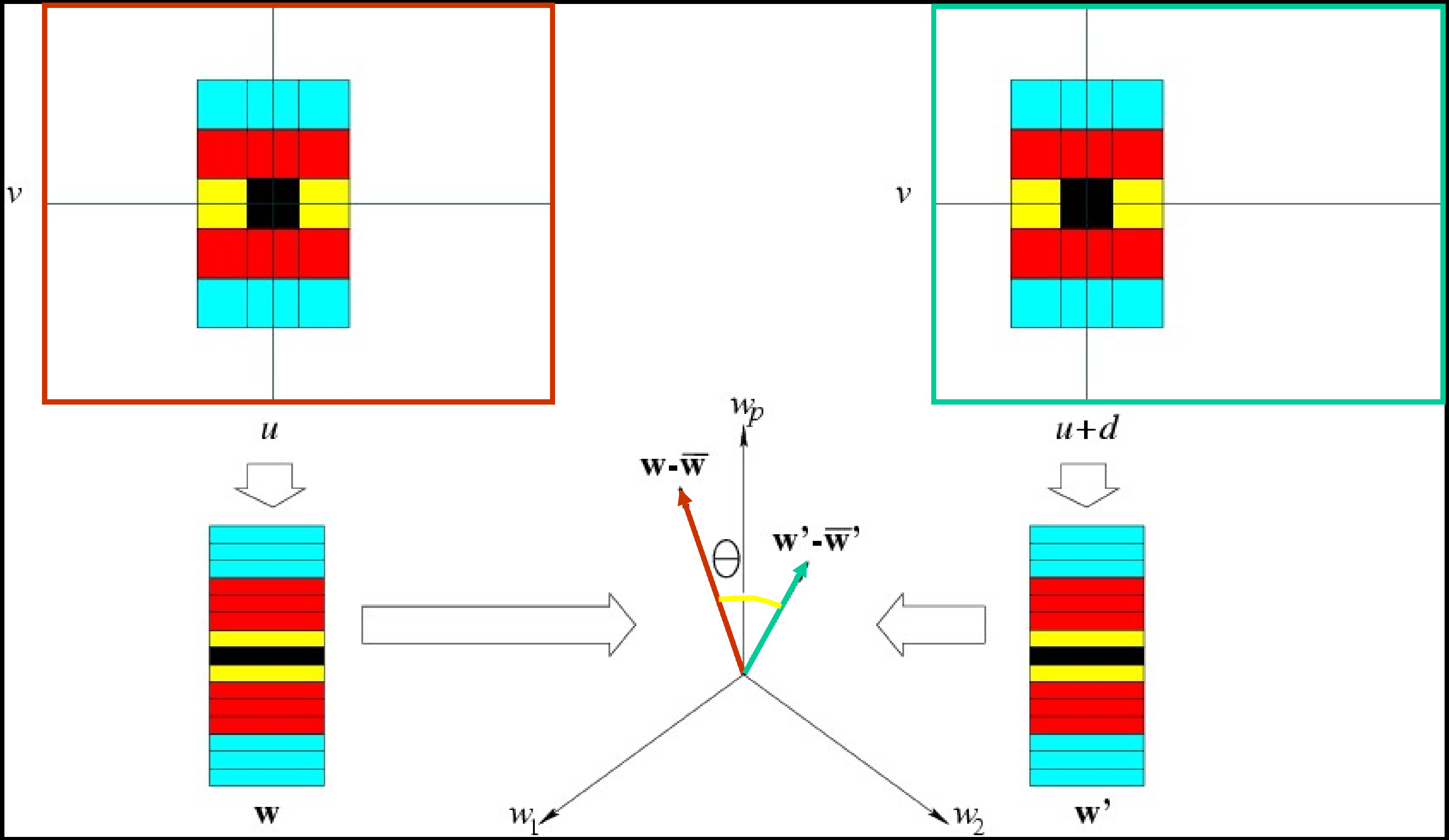
Iterate: $C = \Sigma C_e - w\Sigma C_i + C_0$.


I
1 2 3 4 5 6
1 2 3 4 5 6
u


1 2 3 4 5 6
1 2 3 4 5 6



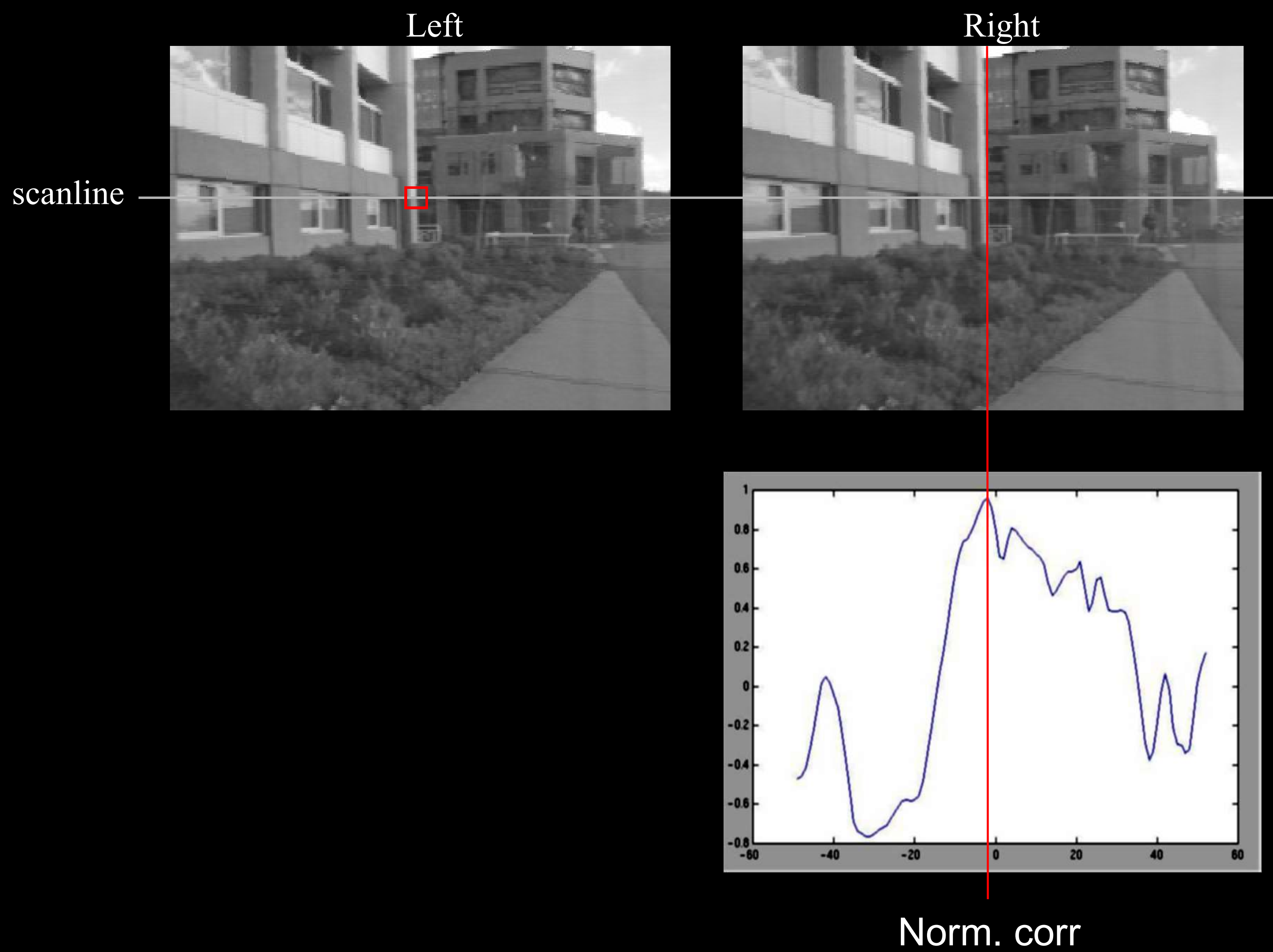
Correlation Methods (1970--)



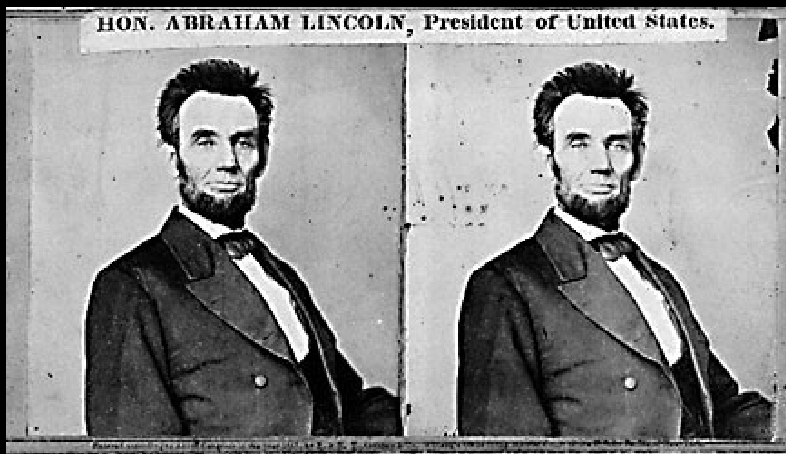
Slide the window along the epipolar line until *w.w'* is maximized.

Normalized Correlation: minimize θ instead. ⟺ Minimize $|w-w'|^2$.

Correlation-based methods



Failures of correlation-based methods

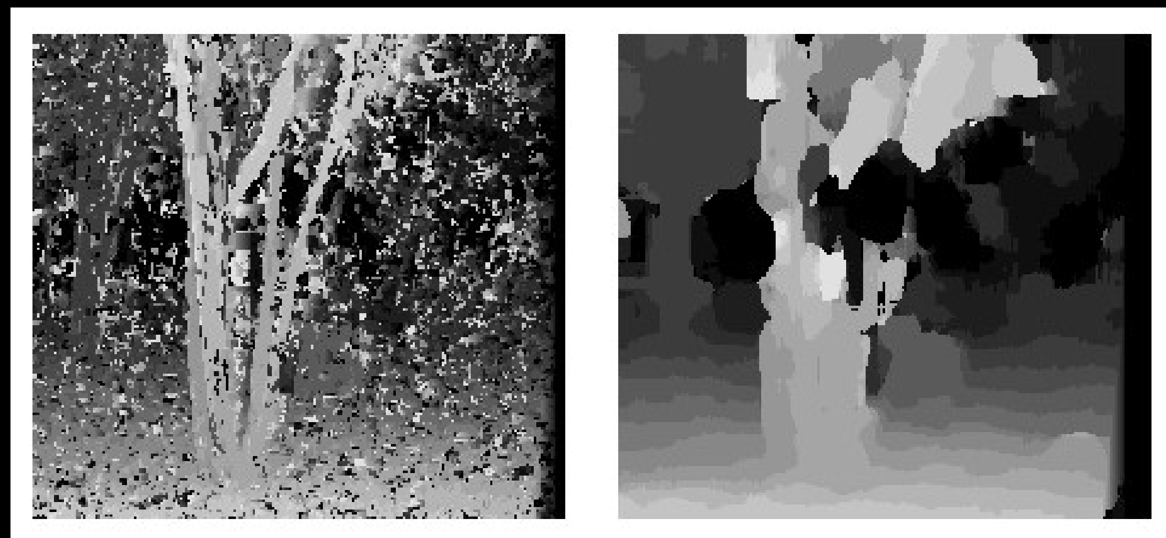


Textureless surfaces

Occlusions, repetition

Non-Lambertian surfaces, specularities

Effect of window size



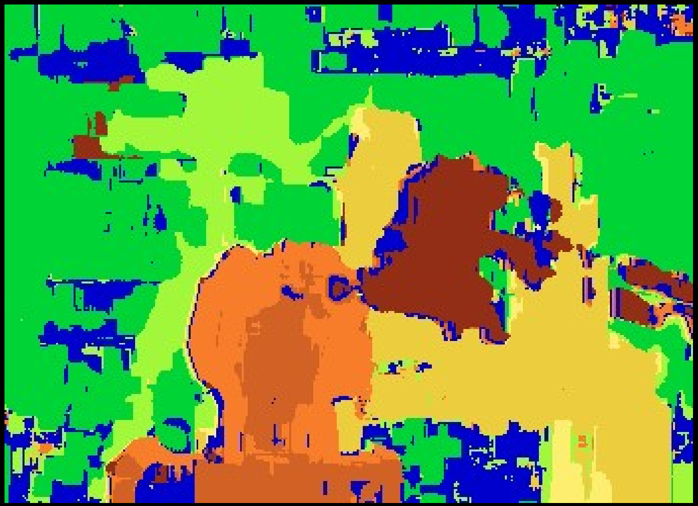
W = 3 W = 20

- Smaller window
  - + More detail
  - • More noise
- Larger window
  - + Smoother disparity maps
  - • Less detail

Results

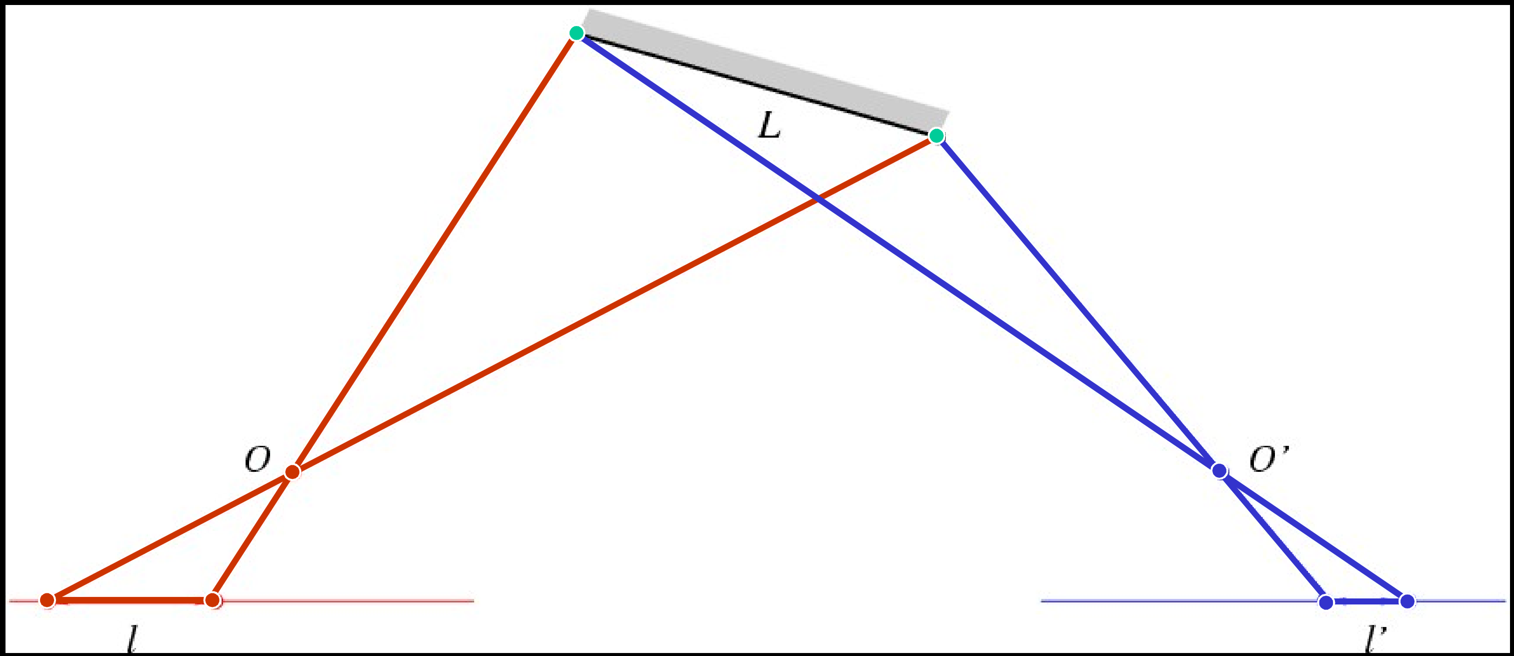
Data

Correlation-based matching

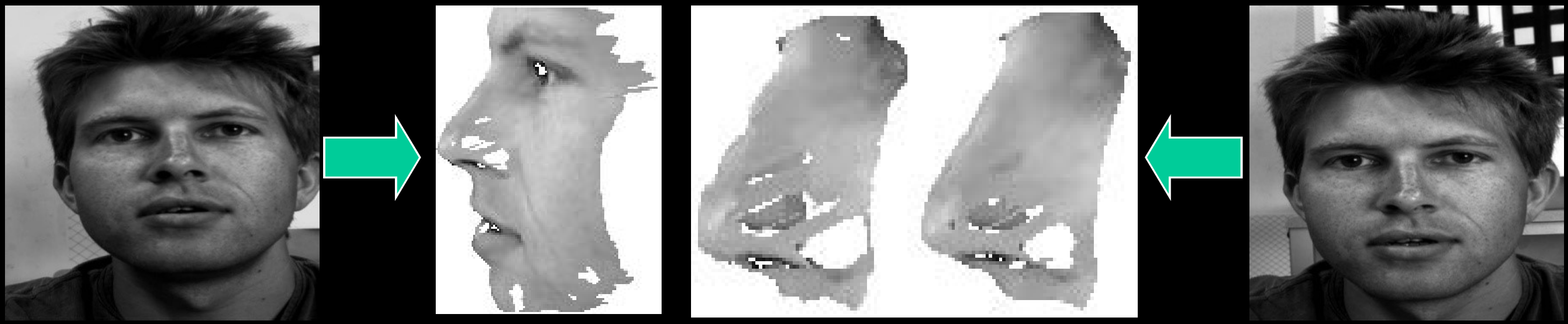


Ground truth

Correlation Methods: Foreshortening Problems



Solution: add a second pass using disparity estimates to warp the correlation windows, e.g. Devernay and Faugeras (1994).



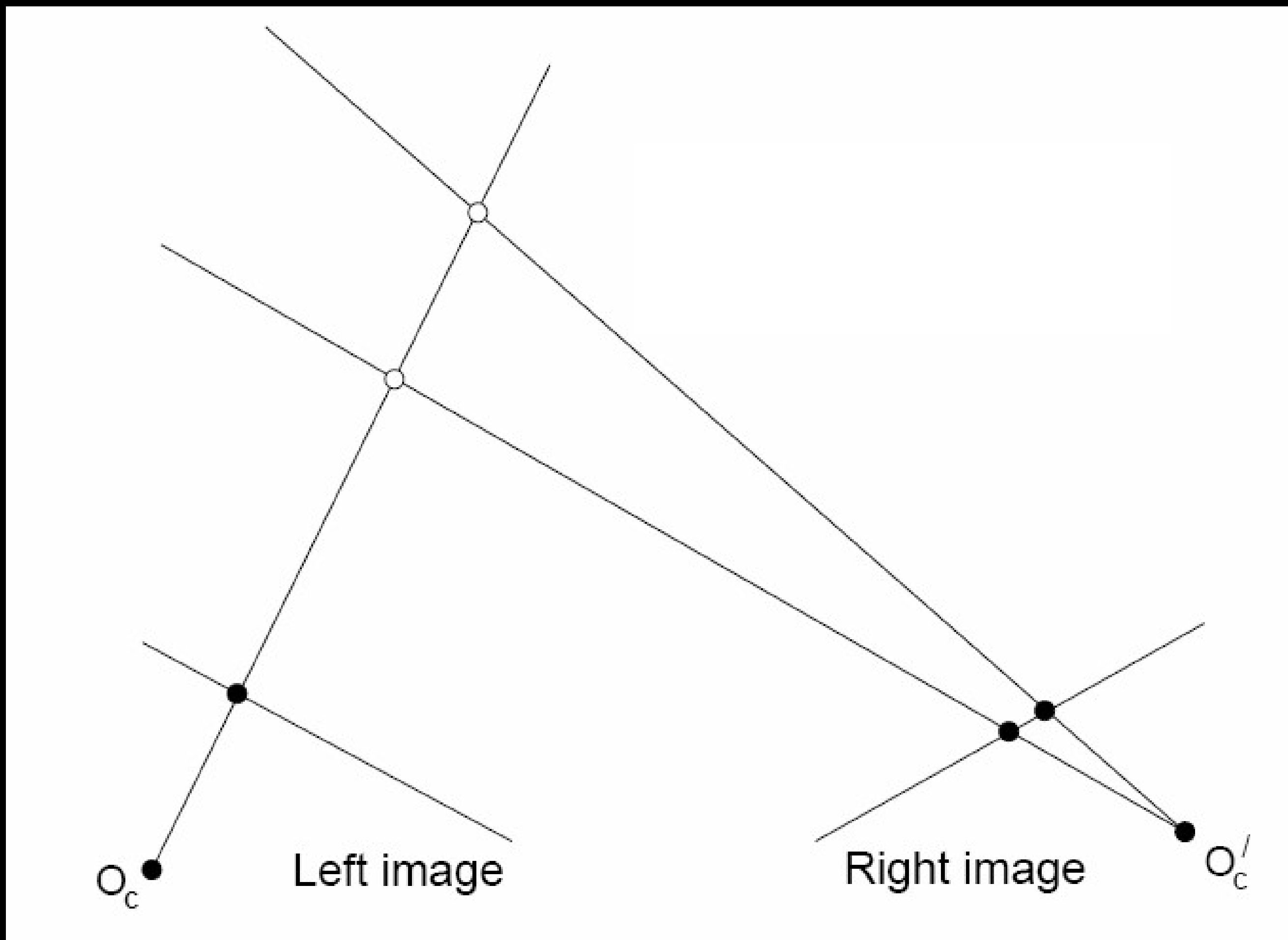
Reprinted from "Computing Differential Properties of 3D Shapes from Stereopsis without 3D Models," by F. Devernay and O. Faugeras, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1994).

How can we improve window-based matching?

- The similarity constraint is local: each reference window is matched independently.
- Need to enforce global correspondence constraints.

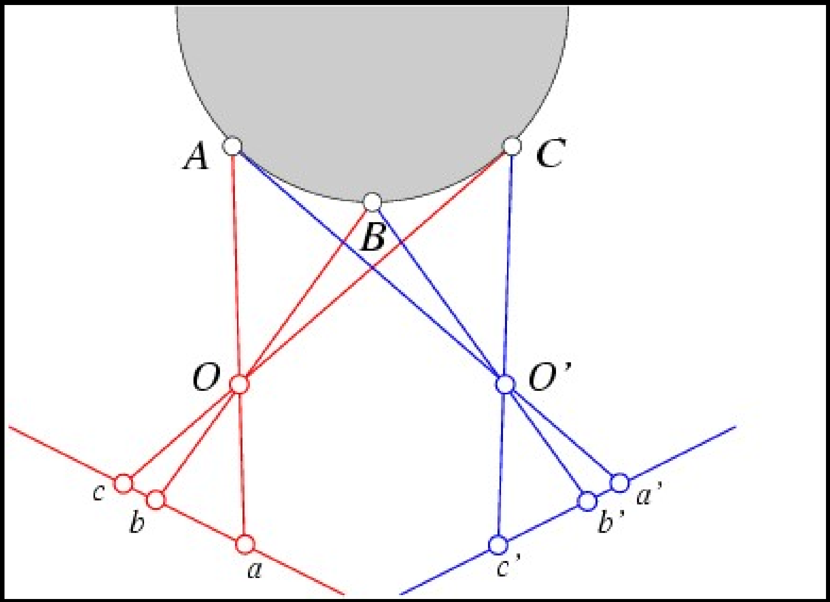
Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image



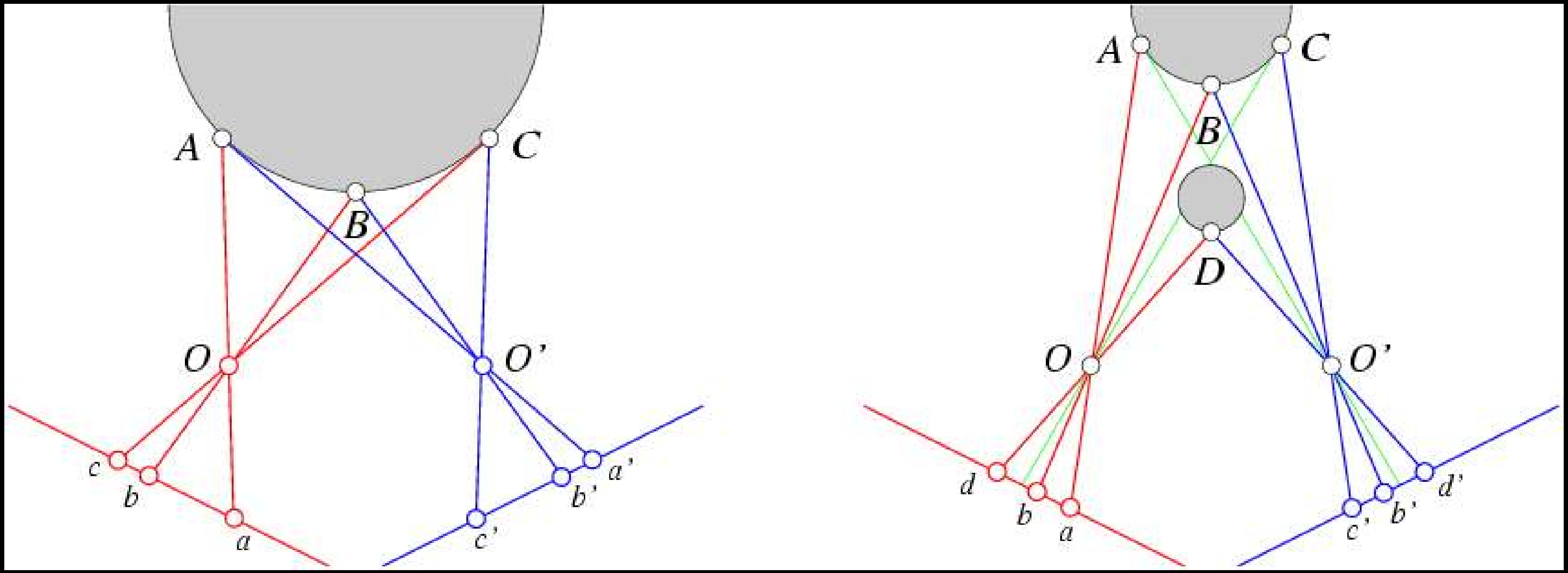
Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views



Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views



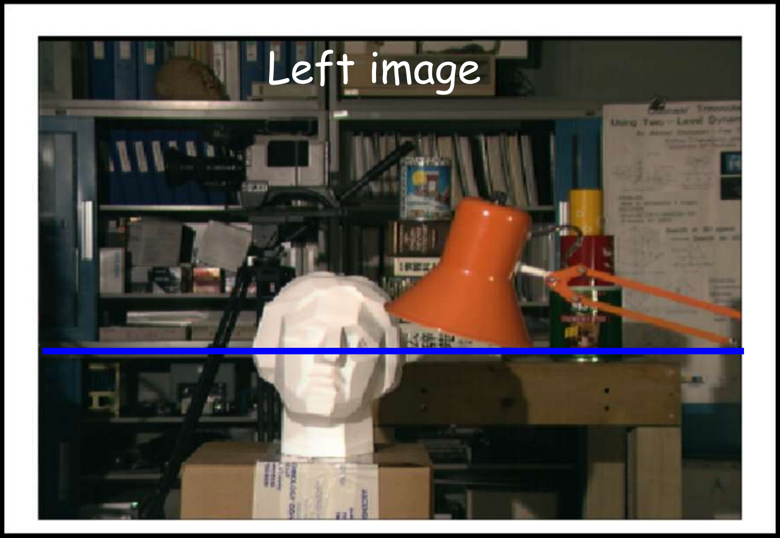
Ordering constraint does not (always) hold

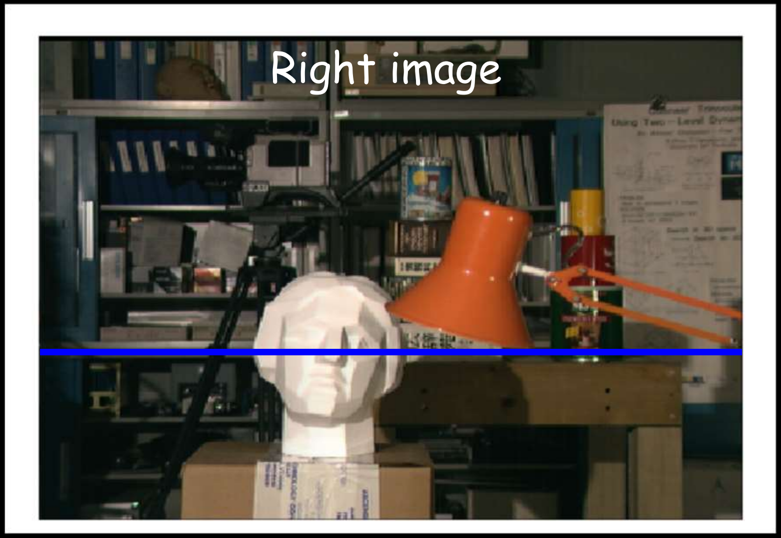
Non-local constraints

- Uniqueness
  - For any point in one image, there should be at most one matching point in the other image
- Ordering
  - Corresponding points should be in the same order in both views
- Smoothness
  - We expect disparity values to change slowly (for the most part)

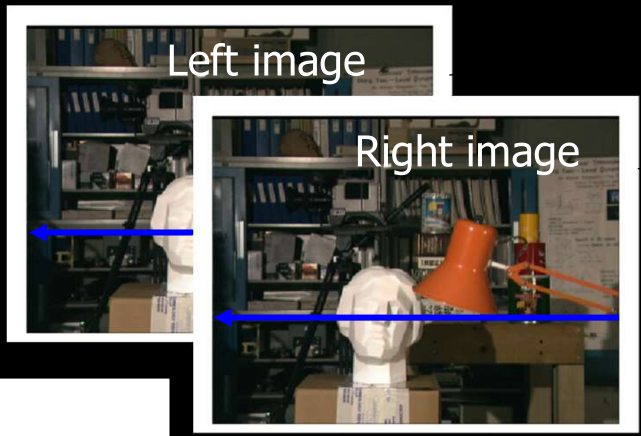
Scanline stereo

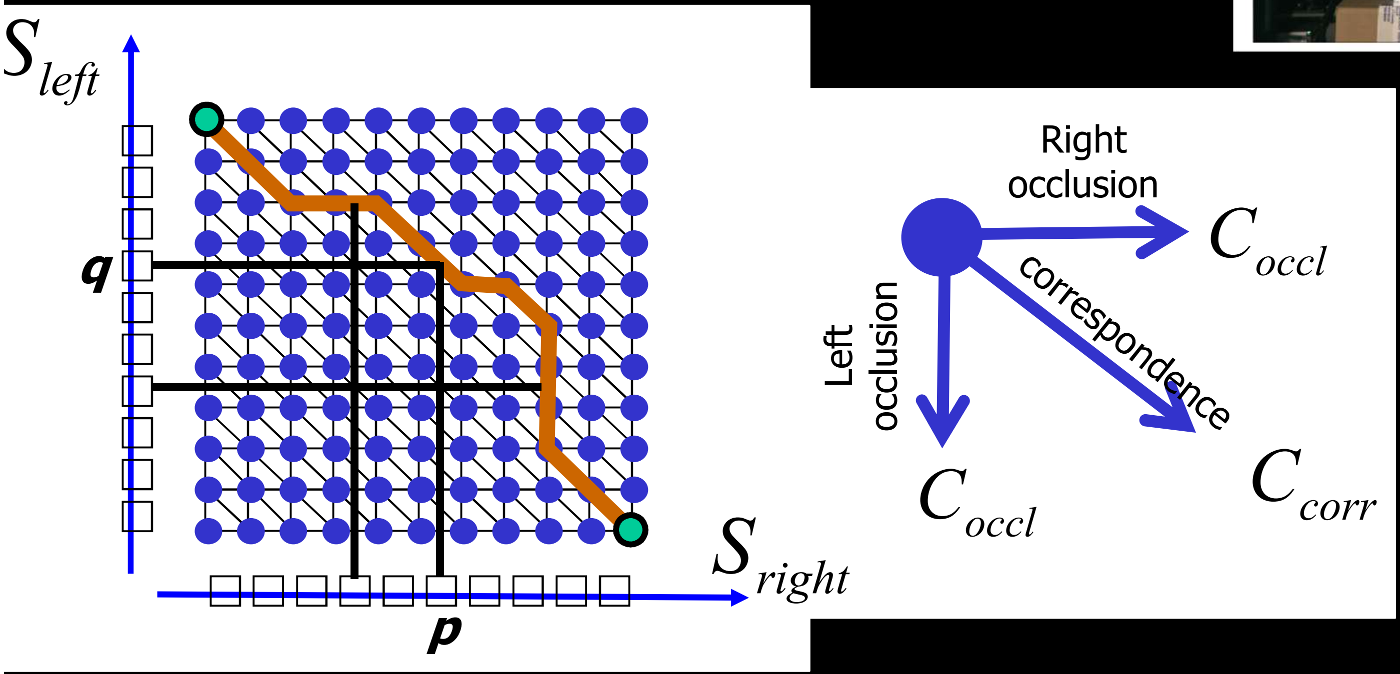
- Try to coherently match pixels on the entire scanline
- Different scanlines are still optimized independently





"Shortest paths" for scan-line stereo



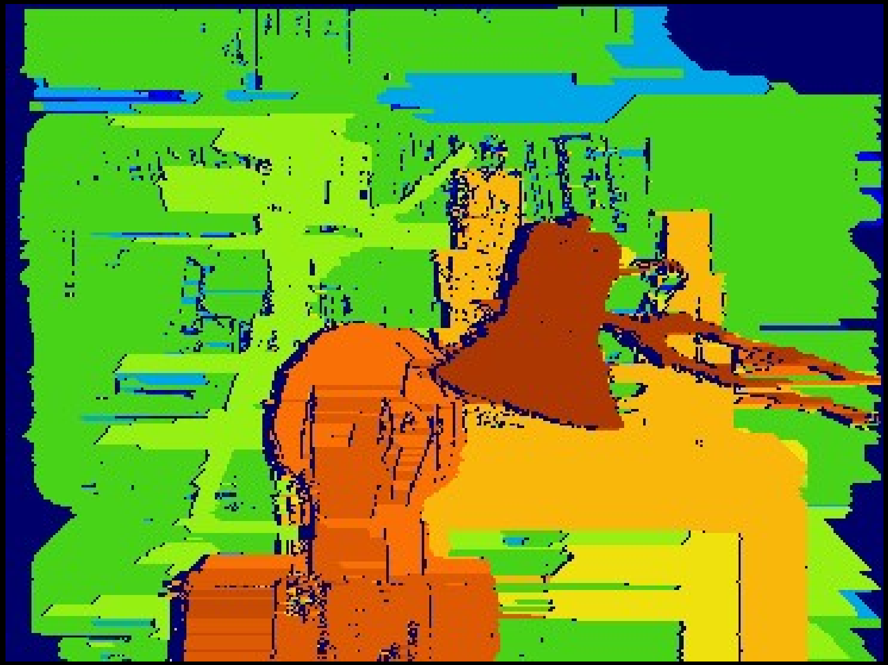


Can be implemented with dynamic programming
(Baker & Binford'81, Ohta & Kanade '85)

Slide credit: Y. Boykov

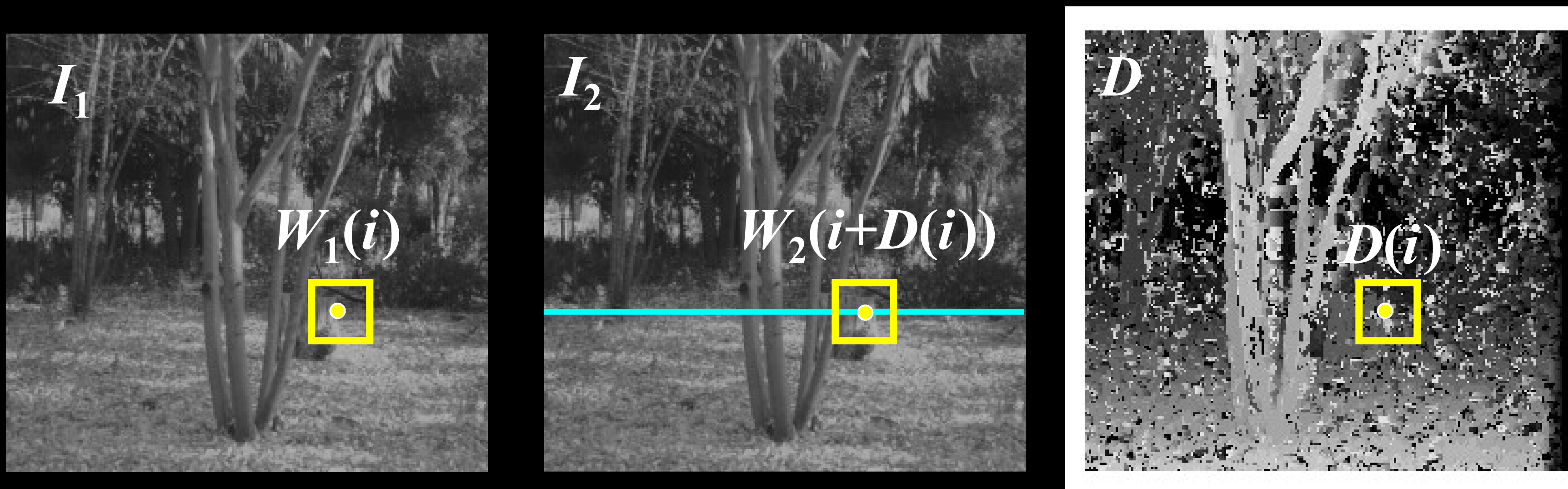
Shortest path stereo in real life

- Scanline stereo generates streaking artifacts



- Can't use dynamic programming to find spatially coherent disparities and correspondences on a 2D grid

Stereo matching as energy minimization



$$E = \alpha\, E_{\text{data}}(I_1, I_2, D) + \beta\, E_{\text{smooth}}(D)$$

$$E_{\text{data}} = \sum_i \big(W_1(i) - W_2(i + D(i))\big)^2 \qquad E_{\text{smooth}} = \sum_{\text{neighbors } i,j} \rho\big(D(i) - D(j)\big)$$

- Energy functions of this form can be minimized using "graph cuts" (aka min-cut/max-flow algorithms)

Y. Boykov, O. Veksler, and R. Zabih, Fast Approximate Energy Minimization via Graph Cuts, PAMI 2001

Combinatorial optimization with unary and binary terms

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} E_i(x_i) + \sum_{1 \leq i < j \leq n} E_{ij}(x_i, x_j)$$

Binary variables

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i x_i + \sum_{1 \leq i < j \leq n} \beta_{ij} x_i x_j$$

Quadratic pseudo-Boolean
function optimization

Generalization to integer variables

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} E_i(x_i) + \sum_{1 \le i < j \le n} E_{ij}(x_i, x_j)$$

- n integer variables in 0..K-1

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i x_i + \sum_{1 \le i < j \le n} \beta_{ij} x_i x_j$$

- nK binary variables (Darbon, 2009)
- x^k=0 if $x \le k$ and 1 otherwise

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{k=0}^{K-1} \alpha_i^k x_i^k + \sum_{1 \le i < j \le n} \sum_{k,l=0}^{K-1} \beta_{ij}^{kl} x_i^k x_j^l$$

Quadratic integer function optimization

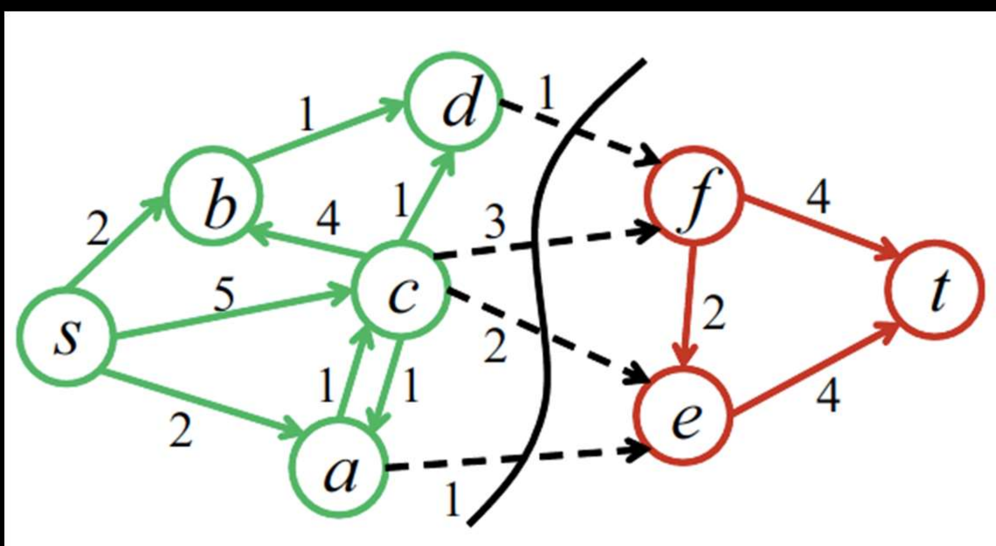
$$E(\boldsymbol{x}) = \sum_{i=1}^{n} E_i(x_i) + \sum_{1 \le i < j \le n} E_{ij}(x_i, x_j)$$

Submodular case

$$\beta_{ij} \le 0$$
$$\beta_{ij}^{kl} \le 0$$

Otherwise
NP hard

Min-cut max-flow problems
(Boros & Hammer, 2002)



Efficient exact algorithms
(Ford & Fulkerson '56)
(Goldberg & Tarjan '88)
(Boykov & Kolgomorov '04)

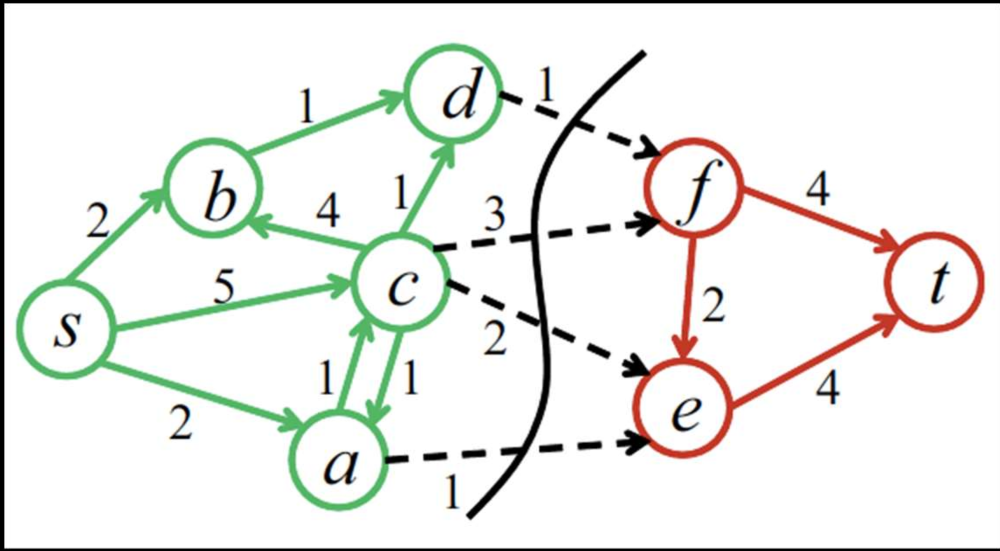
Quadratic integer function optimization

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} E_i(x_i) + \sum_{1 \le i < j \le n} E_{ij}(x_i, x_j)$$

For example: $E_{ij}(x_i, x_j) = g(x_i - x_j)$ with g convex (Ishikawa '03)

Otherwise NP hard

Min-cut max-flow problems (Boros & Hammer, 2002)



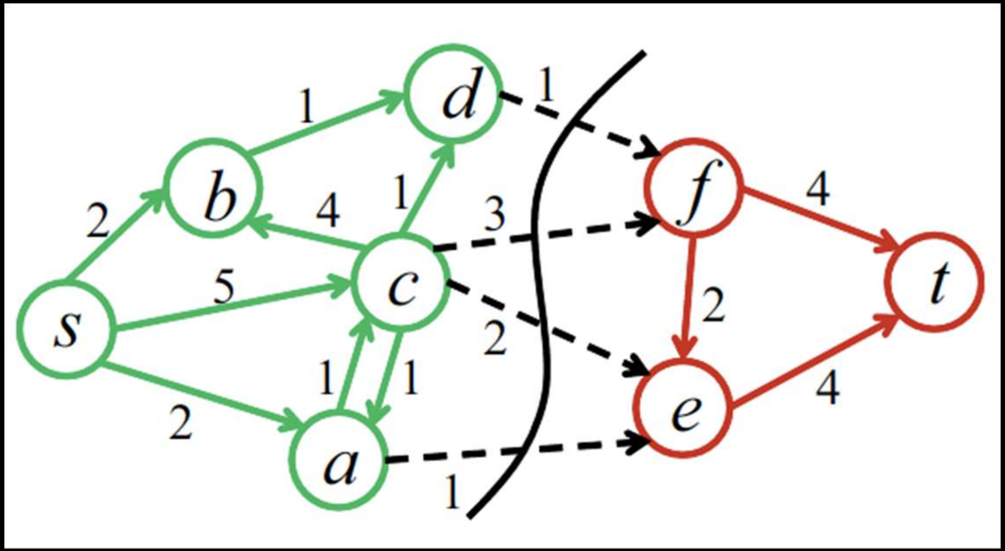
Efficient exact algorithms
(Ford & Fulkerson '56)
(Goldberg & Tarjan '88)
(Boykov & Kolgomorov '04)

Quadratic pseudo-Boolean function optimization

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} E_i(x_i) + \sum_{1 \leq i < j \leq n} E_{ij}(x_i, x_j)$$

For example: $E_{ij}(x_i, x_j) = g(x_i - x_j)$ with g convex (Ishikawa '03)

Min-cut max-flow problems (Boros & Hammer, 2002)



Otherwise NP hard

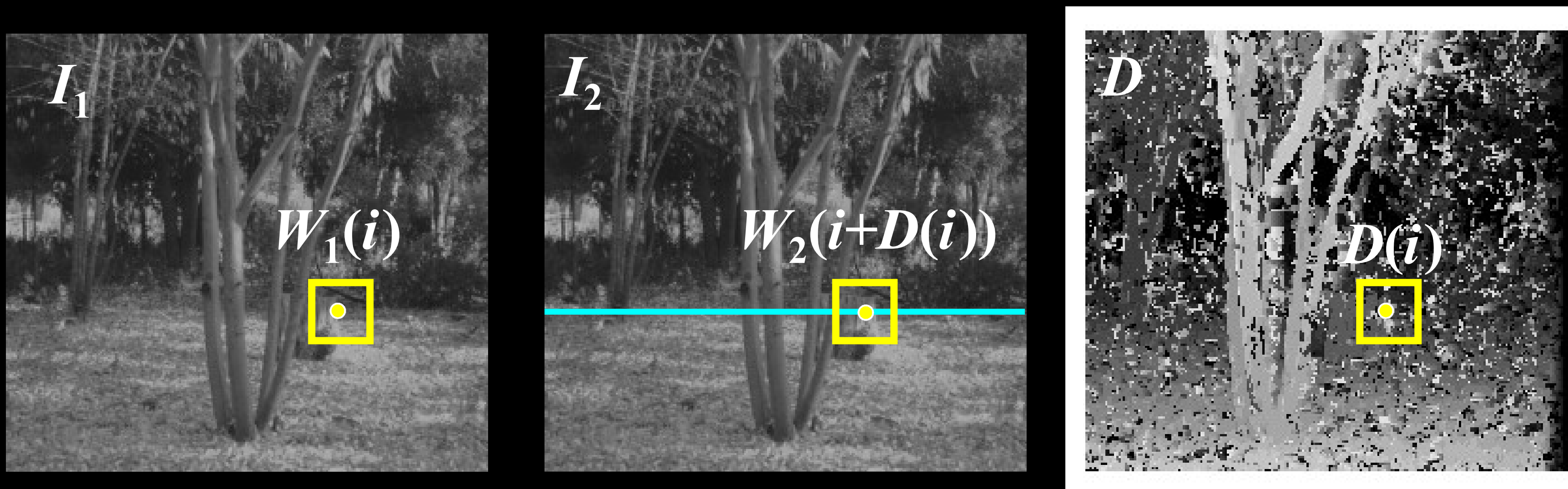
Efficient approximate algorithms (Boykov et al.'01)

Combinatorial optimization:

- Submodularity is "too restrictive" for certain stereo settings (use non-convex g for example)

- Use iterative approximate solutions such as alpha expansion (Boykov et al. 2001)

Back to Stereopsis as energy minimization..



$$E = \alpha\, E_{\text{data}}(I_1, I_2, D) + \beta\, E_{\text{smooth}}(D)$$

$$E_{\text{data}} = \sum_i \left(W_1(i) - W_2(i + D(i))\right)^2 \qquad E_{\text{smooth}} = \sum_{\text{neighbors } i,j} \rho\left(D(i) - D(j)\right)$$

- Energy functions of this form can be minimized using "graph cuts" (aka min-cut/max-flow algorithms)

Y. Boykov, O. Veksler, and R. Zabih, Fast Approximate Energy Minimization via Graph Cuts, PAMI 2001

Stereo matching as energy minimization

- Note: the above formulation does not treat the two images symmetrically, does not enforce uniqueness, and does not take occlusions into account
- It is possible to come up with an energy that does all these things, but it is a bit more complex
  - Defined over all possible sets of matches, not over all disparity maps with respect to the first image
  - Includes an *occlusion term*
  - The smoothness term looks different and more complicated

V. Kolmogorov and R. Zabih, "Computing Visual Correspondences with Occlusions using Graph Cuts, ICCV 2001

Results

Graph cuts

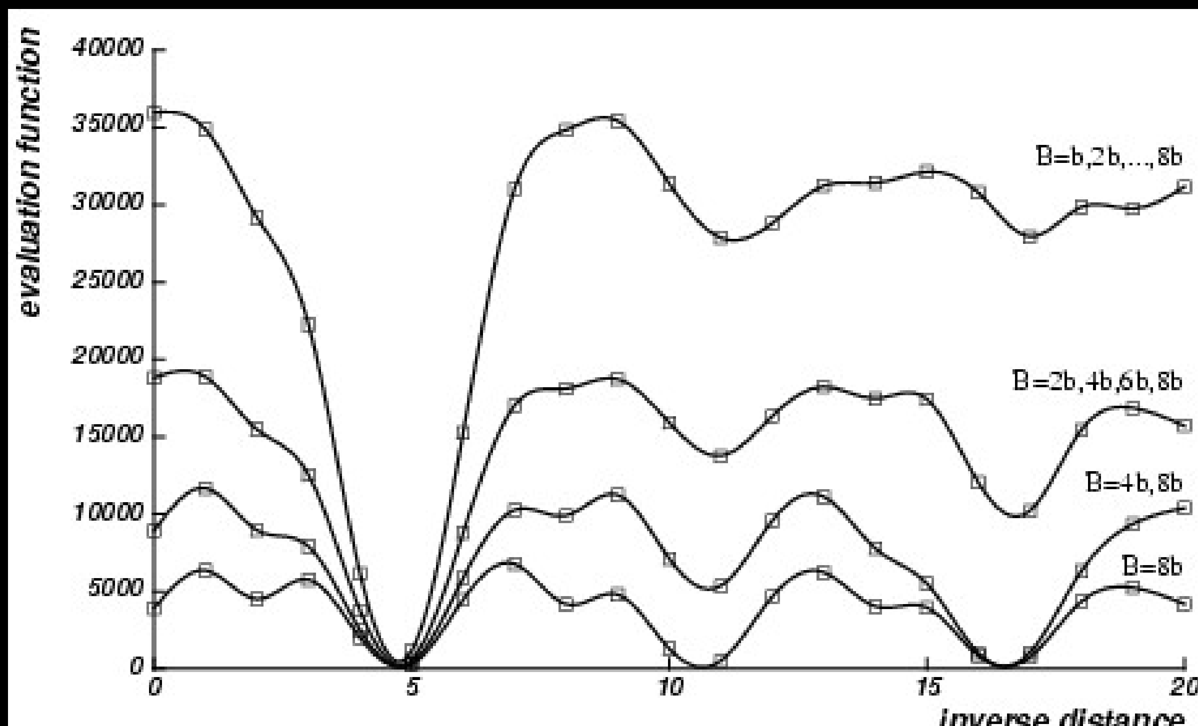
Ground truth

Y. Boykov, O. Veksler and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts, PAMI 2001.

For the latest and greatest: http://www.middlebury.edu/stereo

More Views (Okutami and Kanade, 1993)

Pick a reference image, and slide the corresponding window along the corresponding epipolar lines of all other images, using inverse depth relative to the first image as the search parameter.



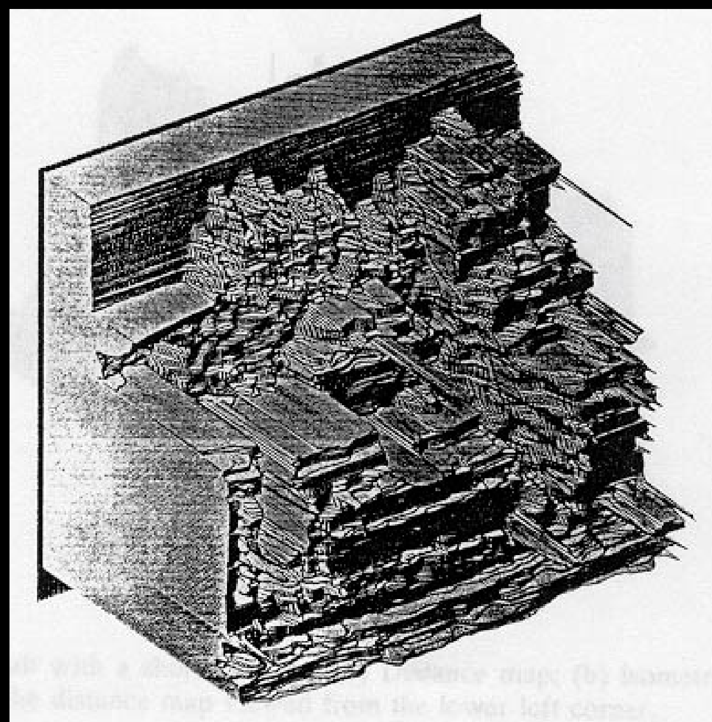
Reprinted from “A Multiple-Baseline Stereo System,” by M. Okutami and T. Kanade, IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(4):353-363 (1993). \copyright 1993 IEEE.

Use the sum of correlation scores to rank matches.

I1

I2

I10



Reprinted from “A Multiple-Baseline Stereo System,” by M. Okutami and T. Kanade, IEEE Trans. on Pattern Analysis and Machine Intelligence, 15(4):353-363 (1993). \copyright 1993 IEEE.

Multi-view geometry questions

- Scene geometry (structure): Given 2D point matches in two or more images, where are the corresponding points in 3D?
- Correspondence (fusion): Given a point in just one image, how does it constrain the position of the corresponding point in another image?
- Camera geometry (motion): Given a set of corresponding points in two or more images, what are the camera matrices for these views?

The Euclidean (perspective) Structure-from-Motion Problem

Given m (internally) calibrated perspective images of n fixed points $\boldsymbol{P}_j$ we can write

$$\begin{cases} u_{ij} = \dfrac{\boldsymbol{m}_{i1} \cdot \boldsymbol{P}_j}{\boldsymbol{m}_{i3} \cdot \boldsymbol{P}_j} \\ v_{ij} = \dfrac{\boldsymbol{m}_{i2} \cdot \boldsymbol{P}_j}{\boldsymbol{m}_{i3} \cdot \boldsymbol{P}_j} \end{cases} \quad \text{for} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, \ldots, n.$$

Problem: estimate the m 3x4 matrices $\mathcal{M}_i = [\mathrm{R}_i \; t_\mathrm{i}]$ and the n positions $\boldsymbol{P}_\mathrm{j}$ from the mn correspondences $\boldsymbol{p}_\mathrm{ij}$.

$2mn$ equations in $11m$ (or rather $5m$)+$3n$ unknowns

Overconstrained problem, that can be solved using (non-linear) least squares!

The Euclidean Ambiguity of Euclidean SFM

When the intrinsic parameters are known (normalized coordinates)

If R_i, $\boldsymbol{t}_i$, and $\boldsymbol{P}_j$ are solutions,

$$p_{ij} = \frac{1}{z_{ij}} \left(\begin{bmatrix} \mathcal{R}_i & \boldsymbol{t}_i \end{bmatrix} \begin{bmatrix} \mathcal{R} & \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \right) \left(\begin{bmatrix} \mathcal{R}^T & -\mathcal{R}^T \boldsymbol{t} \\ \boldsymbol{0}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{P}_j \\ 1 \end{bmatrix} \right) = \frac{1}{z_{ij}} \begin{bmatrix} \mathcal{R}'_i & \boldsymbol{t}'_i \end{bmatrix} \begin{bmatrix} \boldsymbol{P}'_j \\ 1 \end{bmatrix}$$

So are R_i', $\boldsymbol{t}_i'$, and $\boldsymbol{P}_j'$, where

$$\mathcal{R}'_i = \mathcal{R}_i \mathcal{R},\ \boldsymbol{t}'_i = \mathcal{R}_i \boldsymbol{t} + \boldsymbol{t}_i,\ \text{and}\ \boldsymbol{P}'_j = \mathcal{R}^T (\boldsymbol{P}_j - \boldsymbol{t}).$$

In fact, the absolute scale cannot be recovered since:

$$p_{ij} = \frac{1}{\lambda z_{ij}} \begin{bmatrix} \mathcal{R}_i & \lambda \boldsymbol{t}_i \end{bmatrix} \begin{bmatrix} \lambda \boldsymbol{P}_j \\ 1 \end{bmatrix} = \frac{1}{z'_{ij}} \begin{bmatrix} \mathcal{R}_i & \boldsymbol{t}'_i \end{bmatrix} \begin{bmatrix} \boldsymbol{P}'_j \\ 1 \end{bmatrix}$$

Euclidean ambiguity up to a similarity transformation.

Euclidean motion from E (Longuet-Higgins, 1981)

- Given F computed from n > 7 point correspondences, and its SVD $F = UWV^T$, compute $E = U\,\mathrm{diag}(1,1,0)\,V^T$.

- There are two solutions $t' = u_3$ and $t'' = -t'$ to $E^T t = 0$.

- Define
  $R' = UWV^T$ and $R'' = UW^TV^T$ where

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(It is easy to check R' and R'' are rotations.)

- Then $[t_x']R' = -E$ and $[t_x']R'' = E$. Similar reasoning for t''.

- Four solutions. Only two of them place the reconstructed points in front of the cameras.

Singular Value Decomposition

Let $\mathcal{A}$ be an $m \times n$ matrix, with $m \geq n$, then $\mathcal{A}$ can always be written as

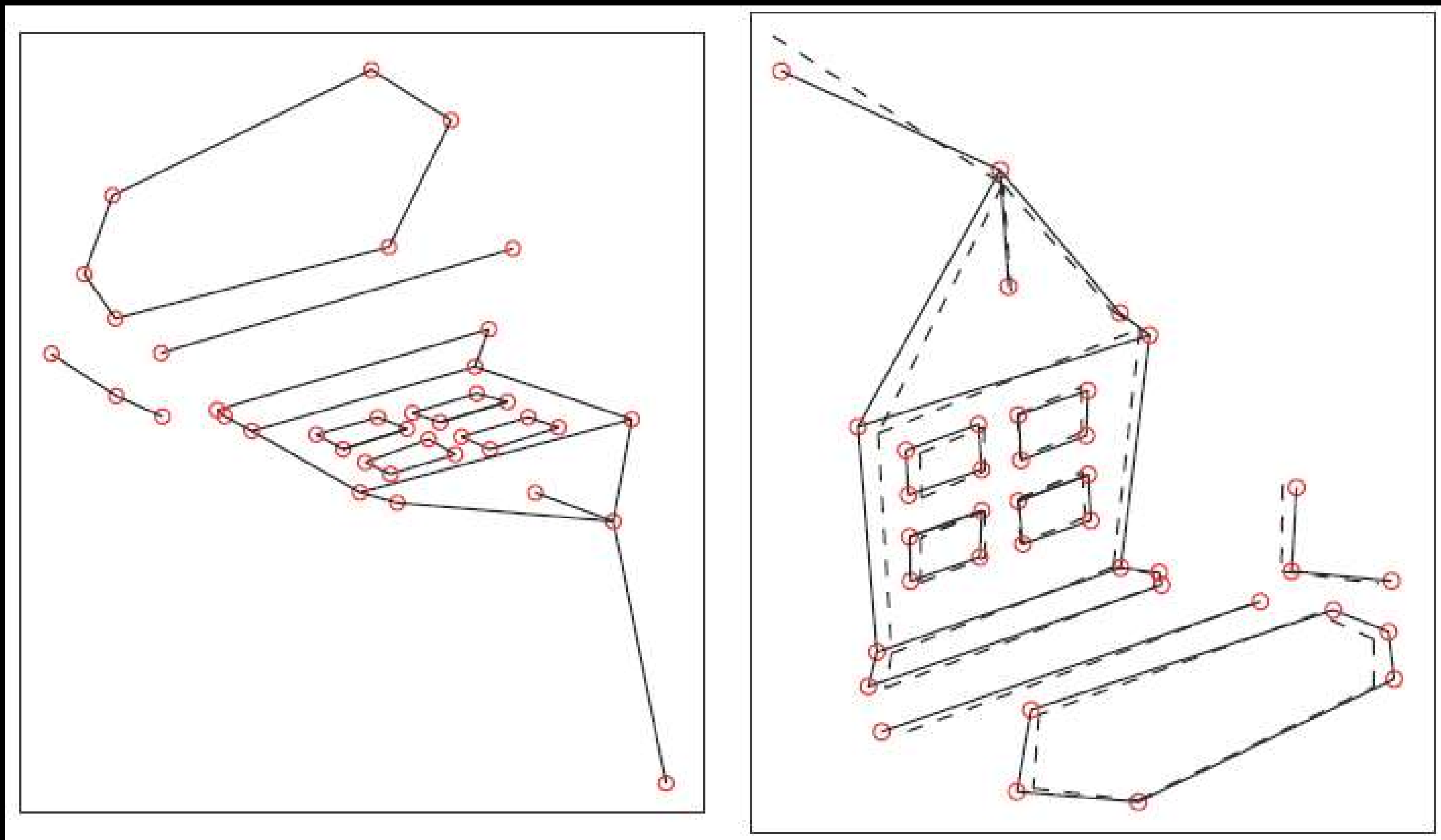
$$\mathcal{A} = \mathcal{U}\mathcal{W}\mathcal{V}^T,$$

where:

- $\mathcal{U}$ is an $m \times n$ column-orthogonal matrix, i.e., $\mathcal{U}^T\mathcal{U} = \mathrm{Id}_m$,
- $\mathcal{W}$ is a diagonal matrix whose diagonal entries w_i $(i = 1, \ldots, n)$ are the singular values of $\mathcal{A}$ with $w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$,
- and $\mathcal{V}$ is an $n \times n$ orthogonal matrix, i.e., $\mathcal{V}^T\mathcal{V} = \mathcal{V}\mathcal{V}^T = \mathrm{Id}_n$.

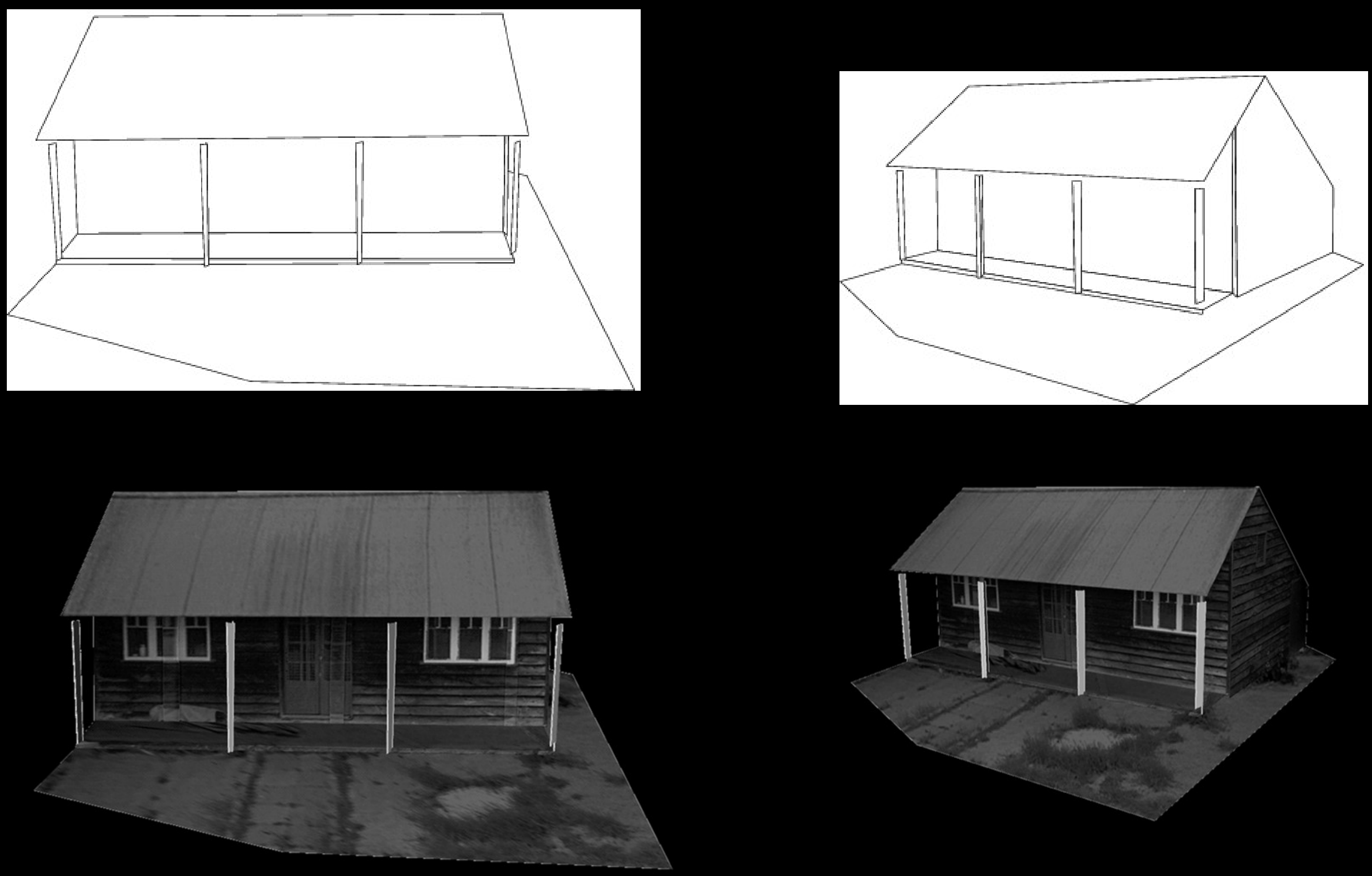
square roots of

Theorem: The singular values of the matrix $\mathcal{A}$ are the eigenvalues of the matrix $\mathcal{A}^T\mathcal{A}$ and the columns of the matrix $\mathcal{V}$ are the corresponding eigenvectors.

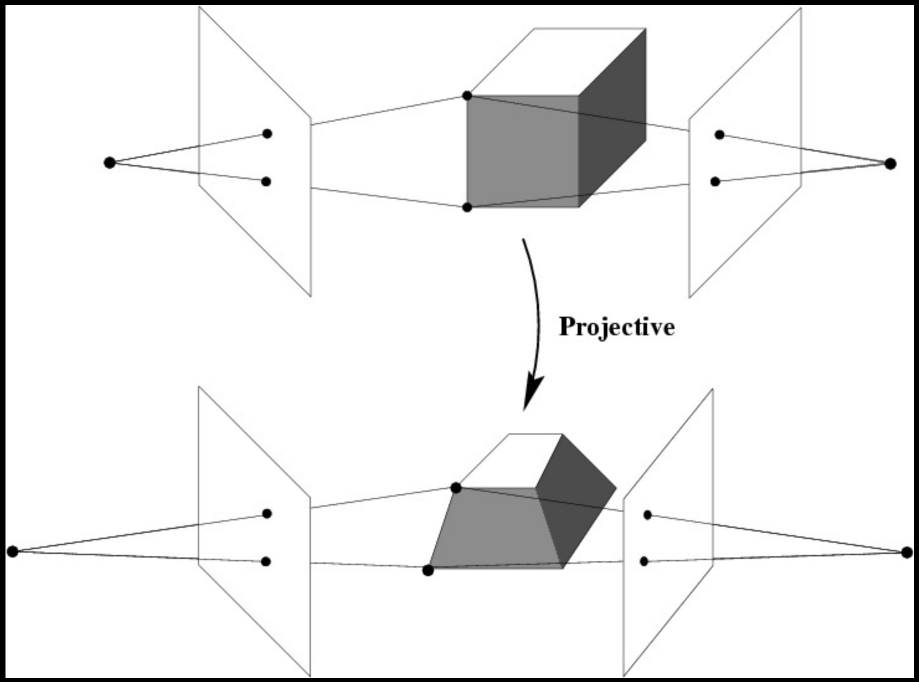


Euclidean reconstruction. Mean relative error: 3.1%

Euclidean (= similarity) ambiguity

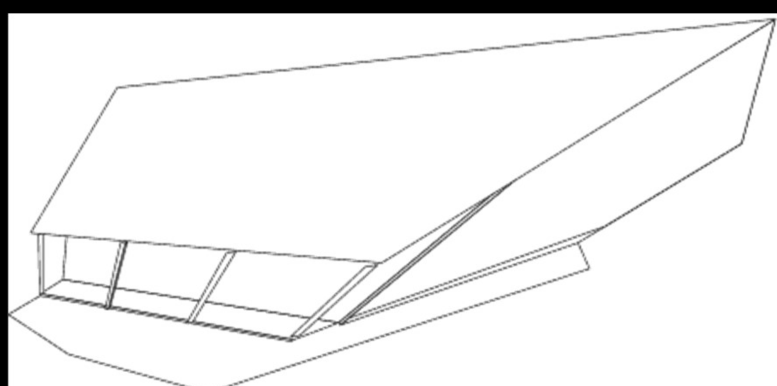
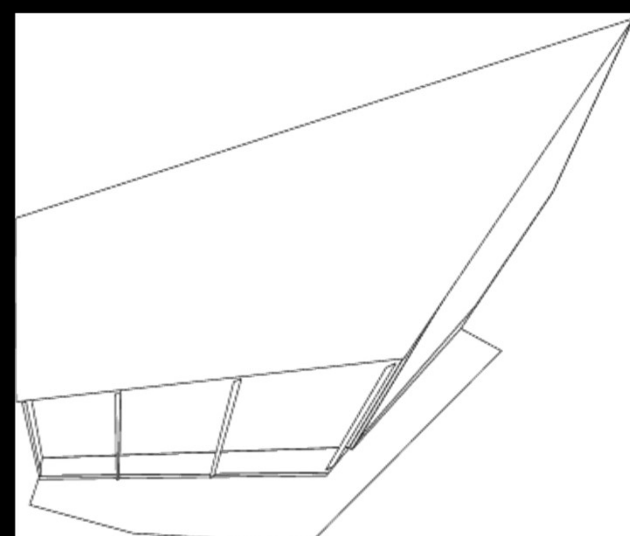


If P is unconstrained: Projective ambiguity



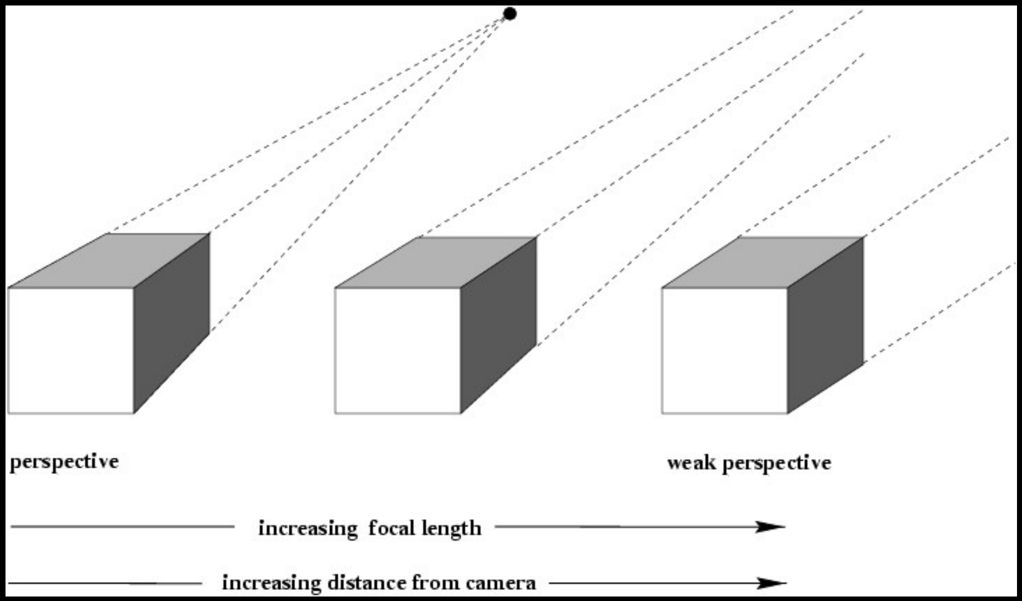
$$\mathbf{x} \approx \mathbf{P}\mathbf{X} = \left(\mathbf{P}\mathbf{Q}_{\mathbf{P}}^{-1}\right)\left(\mathbf{Q}_{\mathbf{P}}\mathbf{X}\right)$$



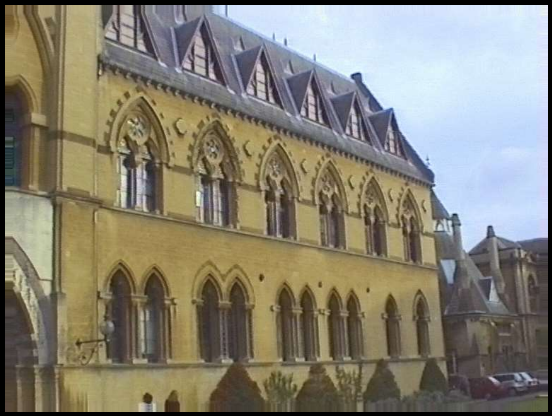



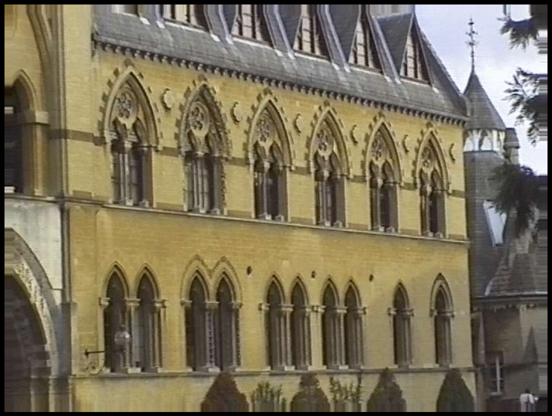
Structure from motion

- Let us now look at simpler, *affine cameras*



center at
infinity





The *Affine* Structure-from-Motion Problem

Given m images of n fixed points P_j we can write

$$\boldsymbol{p}_{ij} = \mathcal{M}_i \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix} = \mathcal{A}_i \boldsymbol{P}_j + \boldsymbol{b}_i \quad \text{for} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, \ldots, n.$$

Problem: estimate the m 2x4 matrices $\mathcal{M}_i$ and the n positions P_j from the mn correspondences $\boldsymbol{p}_{ij}$.

$2mn$ equations in $8m+3n$ unknowns

Overconstrained problem, that can be solved using (non-linear) least squares!

The Affine Ambiguity of Affine SFM

When the intrinsic parameters are unknown

If M_i and P_j are solutions,

$$\boldsymbol{p}_{ij} = \mathcal{M}_i \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix} = (\mathcal{M}_i \mathcal{Q}) \left(\mathcal{Q}^{-1} \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix} \right) = \mathcal{M}'_i \begin{pmatrix} \boldsymbol{P}'_j \\ 1 \end{pmatrix}$$

So are M'_i and P'_j where

$$\mathcal{M}'_i = \mathcal{M}_i \mathcal{Q} \quad \text{and} \quad \begin{pmatrix} \boldsymbol{P}'_j \\ 1 \end{pmatrix} = \mathcal{Q}^{-1} \begin{pmatrix} \boldsymbol{P}_j \\ 1 \end{pmatrix}$$

and

$$\mathcal{Q} = \begin{pmatrix} \mathcal{C} & \boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix} \quad \text{with} \quad \mathcal{Q}^{-1} = \begin{pmatrix} \mathcal{C}^{-1} & -\mathcal{C}^{-1}\boldsymbol{d} \\ \boldsymbol{0}^T & 1 \end{pmatrix}$$

Q is an affine transformation.