

Computer Vision, Lectures 13,14

Professor Hager

<http://www.cs.jhu.edu/~hager>

Outline for Today

- Stereo geometry
- Stereo matching
- Stereo evaluation article

EPIPOLAR GEOMETRY: DERIVATION

$$(\mathbf{P}_1 - \mathbf{T}) \cdot (\mathbf{T} \times \mathbf{P}_1) = 0$$

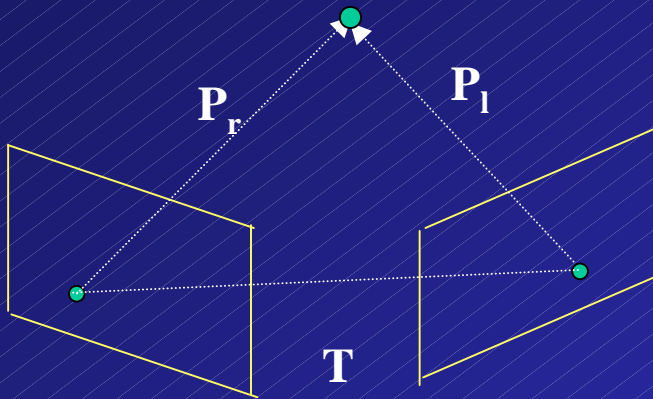
$$\mathbf{P}_r^t \mathbf{R} (\mathbf{T} \times \mathbf{P}_1) = 0$$

$$\mathbf{P}_r^t \mathbf{E} \mathbf{P}_1 = 0$$

where $\mathbf{E} = \mathbf{R} \text{sk}(\mathbf{T})$

$$\text{sk}(\mathbf{T}) = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}$$

The matrix \mathbf{E} is called the *essential matrix* and completely describes the epipolar geometry of the stereo pair



$$\mathbf{P}_r = \mathbf{R}(\mathbf{P}_1 - \mathbf{T})$$

EPIPOLAR GEOMETRY: COMPUTATION

$$\mathbf{p}_r^t \mathbf{E} \mathbf{p}_l = 0 \quad \text{or} \quad \mathbf{r}_r^t \mathbf{F} \mathbf{r}_l = 0$$

Note that, given a correspondence, we can form a linear constraint on E (or F). Both E and F are only unique up to scale, therefore we need $9-1 = 8$ matches, then we can form a system of the form

$$\mathbf{C} \mathbf{e} = 0 \quad \text{where } \mathbf{e} \text{ is the vector of 9 values in E}$$

Using SVD, we can write $\mathbf{C} = \mathbf{U} \mathbf{D} \mathbf{V}^t$

E (or F) is the column of \mathbf{V} corresponding to the least singular value of \mathbf{C} .

WHY?

E (or F) is supposed to be rank deficient; to enforce this, we can compute the SVD of E (or F), set the smallest singular value to 0, then multiply the components to get the corrected F

EPIPOLAR GEOMETRY: RECONSTRUCTION

$$\mathbf{p}_r^t \mathbf{E} \mathbf{p}_l = 0 \quad \text{or} \quad \mathbf{r}_r^t \mathbf{F} \mathbf{r}_l = 0$$

One additional useful fact is that we can use epipolar geometry for reconstruction.

First, note that $\mathbf{E}^t \mathbf{E}$ involves only translation and that $\text{tr}(\mathbf{E}^t \mathbf{E}) = 2 \|\mathbf{T}\|^2$

So, if we normalize by $\sqrt{\text{tr}(\mathbf{E}^t \mathbf{E})/2}$, we compute a new matrix \mathbf{E}' which has unit norm translation \mathbf{T}' up to sign.

We can solve for \mathbf{T}' from \mathbf{E}' (or \mathbf{T} from \mathbf{E} for that matter)

Now define $\mathbf{w}_i = \mathbf{E}'_i \times \mathbf{T}'$ and $\mathbf{R}_i = \mathbf{w}_i + \mathbf{w}_j \times \mathbf{w}_k$

The three values of \mathbf{R}_i for all combinations of 1,2,3 are the rows of the rotation matrix.

EPIPOLAR GEOMETRY: STEREO CORRESPONDENCE

$$\mathbf{p}_r^t \mathbf{E} \mathbf{p}_l = 0 \quad \text{or} \quad \mathbf{r}_r^t \mathbf{F} \mathbf{r}_l = 0$$

One of the important uses of epipolar geometry is that it greatly reduces the complexity of stereo. Given a match in the left image, the appropriate place to look for a match in the right is along the corresponding epipolar line.

Alternatively, it is possible to use epipolar structure to *warp* the image to have parallel epipolar geometry, making stereo search a trivial scan-line search.

Using E to get Nonverged Stereo

- From E we get R and T such that ${}^l p = {}^l R_r {}^r p + {}^l T k$
- Note that T is really the direction we'd like the camera baseline to point in.
- Let $R_x = T$
- Let $R_y = (0,0,1) \times T / |T \times (0,0,1)|$
- Let $R_z = R_x \times R_y$
- Now, $R = [R_x, R_y, R_z]'$ takes point from the left camera to a nonverged camera system, so we have
- ${}^{newl}R = R, {}^{newr}R = R \mid R_r$
 - (note the book uses the transpose of this, i.e. the rotation of the frame rather than the points)

THE FUNDAMENTAL MATRIX AND RECONSTRUCTION

$$\mathbf{p}_r^t \mathbf{E} \mathbf{p}_l = \mathbf{0} \quad \text{or} \quad \mathbf{r}_r^t \mathbf{F} \mathbf{r}_l = \mathbf{0}$$

If we do not know the internal parameters, then the 8 point algorithm can only be used to compute F.

Unfortunately, F has less structure; what we can show is that we can only reconstruct up to a projective transformation (but we won't cover that).

EPIPOLAR GEOMETRY: RECTIFICATION

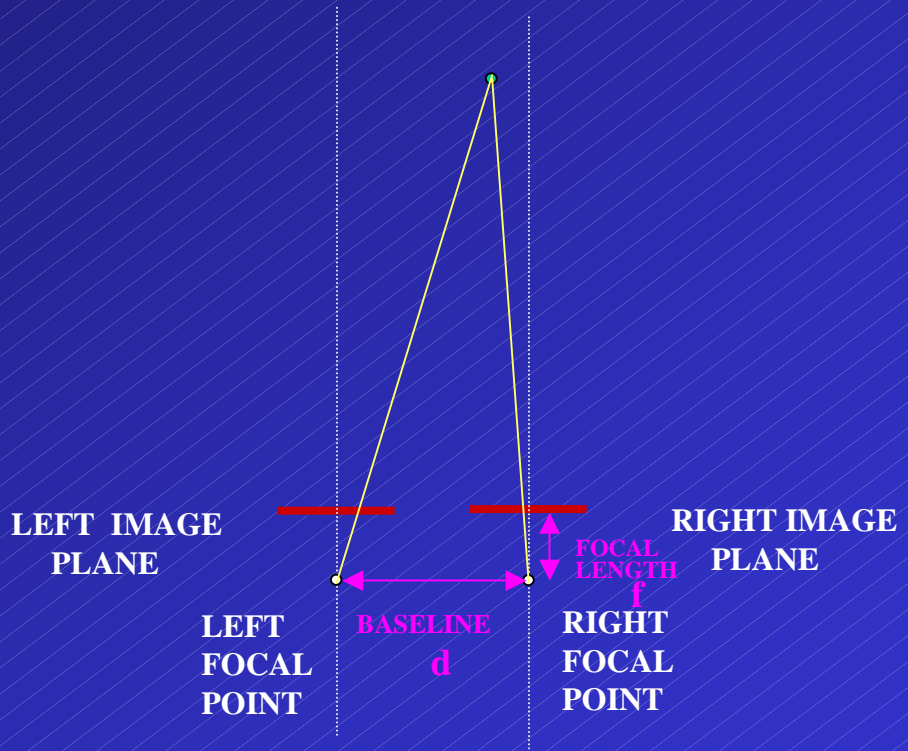
$$\mathbf{p}_r^t \mathbf{E} \mathbf{p}_l = 0 \quad \text{or} \quad \mathbf{r}_r^t \mathbf{F} \mathbf{r}_l = 0$$

Putting all this together, we get the following algorithm:

- 1. Find 8 or more correspondences and compute \mathbf{E} (note we need internal parameters to do this).**
- 2. Given \mathbf{E} , compute \mathbf{T}' and \mathbf{R} .**
- 3. Given \mathbf{T}' and \mathbf{R} , compute rotations into a non-veged system**
- 4. Rectify the images using these rotations**

BINOCULAR STEREO SYSTEM

- **Correspondence Problem** is a key issue for binocular stereo -- namely identify image features in respective images that correspond to exactly the same world object point.
- Clearly localization of image features (e.g., edges) is of critical importance to 3D measurement accuracy.



(2D topdown view)

Computing the Disparity Range

- The first step in correspondence search is to compute the range of disparities to search
 - The *horopter* is the set of distances which have disparity zero (or very close to zero) for a verged system. Human stereo only takes place within the horopter.
- We can assume a non-verged system. Therefore, we have
 - $u_l - u_r = f b/z$
 - substitute $u_r = u_l + \nabla d \rightarrow \delta d = f b/z$
 - given a range z_{\min} to z_{\max} , we calculate
 - $\nabla d_{\min} = f b / z_{\max}$
 - $\nabla d_{\max} = f b / z_{\min}$
- Thus, for each point u_l in the left image, we will search points $u_l + \nabla d_{\min}$ to $u_l + \nabla d_{\max}$ in the right.
- Note we can turn this around and start at a point u_r and search from $u_r - \nabla d_{\max}$ to $u_r - \nabla d_{\min}$

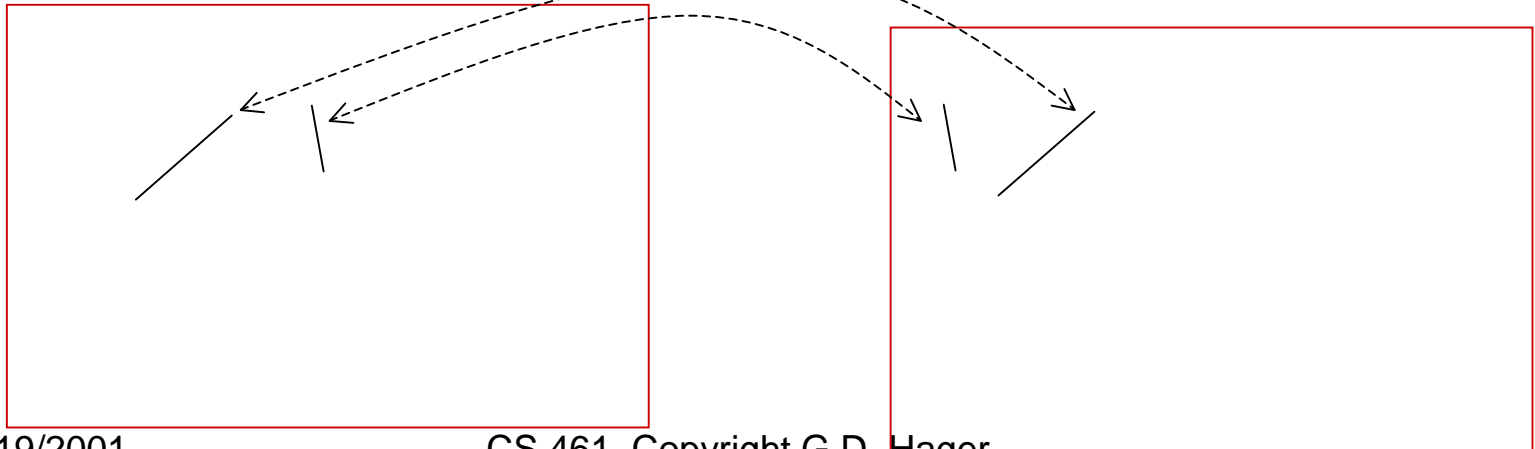
MATCHING AND CORRESPONDENCE

- **Two major approaches**

- feature-based
- region based

In feature-based matching, the idea is to pick a feature type (e.g. edges), define a matching criteria (e.g. orientation and contrast sign), and then look for matches within a disparity range

$$S = \frac{1}{w_0 (l_l - l_r)^2 + w_1 (m_l - m_r)^2 + w_2 (o_l - o_r)^2 + w_3 (c_l - c_r)^2}$$

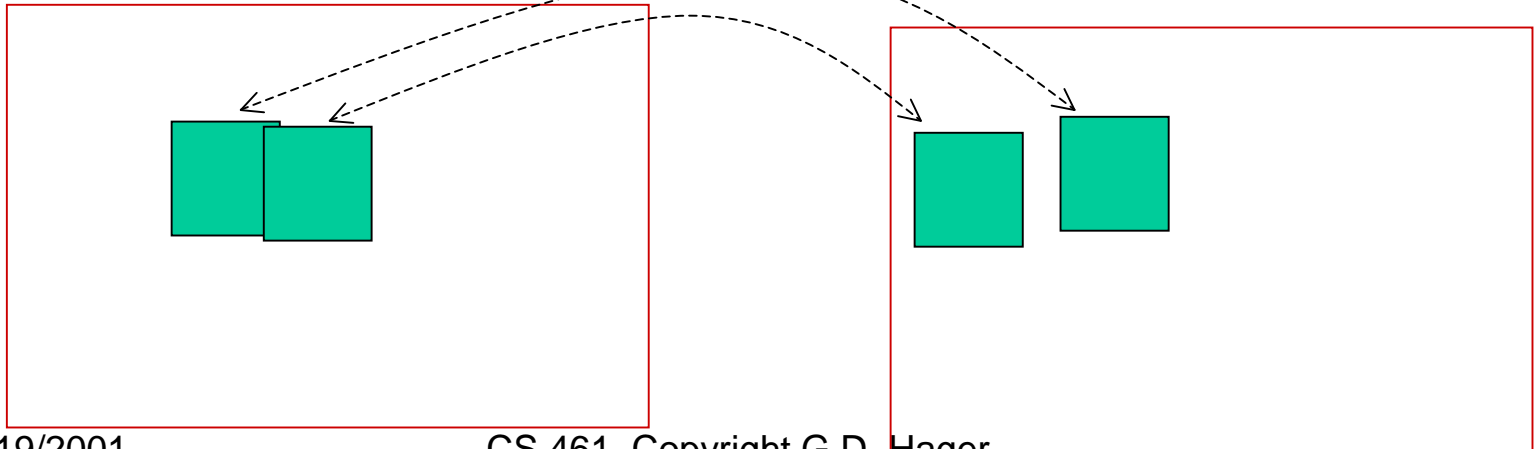


MATCHING AND CORRESPONDENCE

- **Two major approaches**
 - feature-based
 - region based

In region-based matching, the idea is to pick a region in the image and attempt to find the matching region in the second image by maximizing the some measure:

- 1. normalized SSD**
- 2. SAD**
- 3. normalized cross-correlation**



Region Matching

- For each pixel (i,j) of the left image and offset $\nabla i, \nabla j$ in disparity range
 - compute $d(\nabla i, \nabla j) = \sum_{k,l} \psi(I_l(i+k, j+l), I_r(i+k+\nabla i, j+l+\nabla j))$
 - the disparity is the value $(\nabla i, \nabla j)$ that minimizes d
- The result of performing this search over every pixel is the *disparity map*.
- Often, this map is computed at different scales, by performing reduction using a Gaussian pyramid

Matching Metrics

- An obvious solution: minimize the sum of squares
 - think of R and R' as a region and a candidate region in vector form
 - $SSD = \|R - R'\|^2 = \|R\|^2 - 2R \cdot R' + \|R'\|^2$
 - Note that we can change the SSD by making the image brighter or dimmer, or changing contrast
 - As a result, it is common to
 - subtract the mean of both images (removes brightness)
 - normalize by variance (removes contrast)
 - Note taking two derivatives (e.g. a Laplacian) has roughly the same effect!
 - In this case, minimizing SSD is equivalent to maximizing $R \cdot R'$
 - this is the normalized cross correlation!
- Both SSD and NCC are sensitive to outliers
 - $SAD = \sum |R - R'|$ is less sensitive to outliers and thus more robust
 - it is also easier to compute.

Other Matching Metrics

- rank transformation:
 - The value of a window is the # of values less than center pixel
 - Compare values using SAD or SSD on transformed image
 - Invariant over rotations, reflection, and any monotone transformation of gray values
- census transformation
 - A window becomes a bit string based on comparison to center pixel
 - Compare using Hamming distance (# of bits that differ)

- Note both of these are easy to implement in hardware

MATCHING AND CORRESPONDENCE

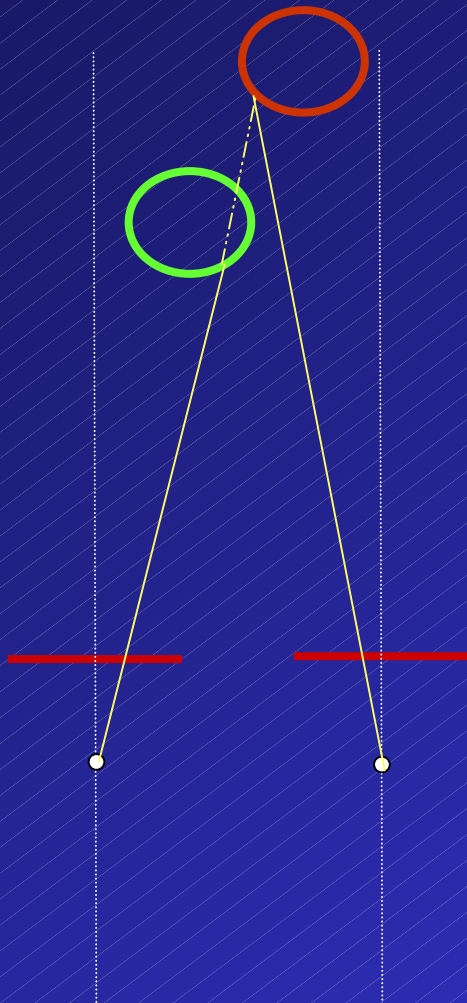
- **Feature-based vs. region-based**
 - feature-based leads to sparse disparity maps
 - interpolation to fill in gaps
 - scale-space approaches to fill in gaps
 - region-based matching only works where there is texture
 - compute a confidence measure for regions
 - apply continuity or match ordering constraints
 - region matching can be sensitive to changes in surface orientation
 - feature-based can be sensitive to feature “drop-outs”

SUMMARY: SIMPLE STEREO

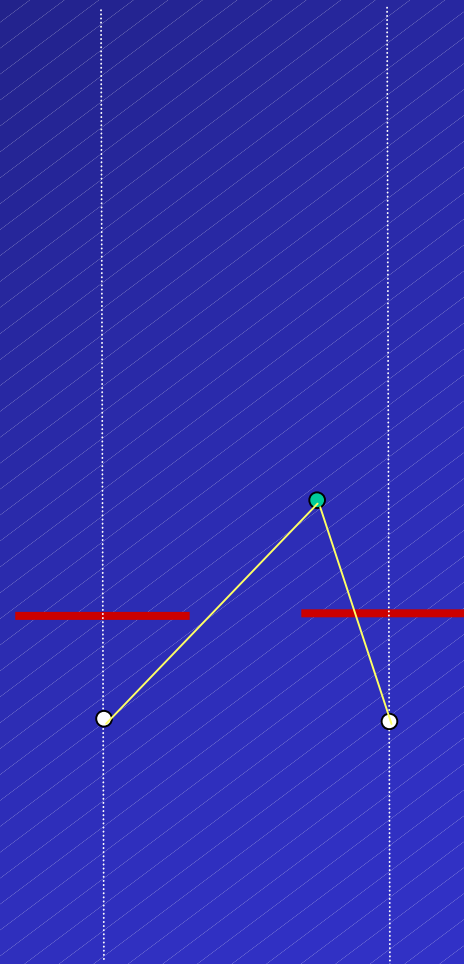
Given two cameras with *known* relative positions in space and known internal parameters:

- 1. Rectify the two images using epipolar geometry.**
- 2. Compute image correspondences using either feature-based or correlation-based matching**
- 3. Convert resulting pixel coordinates to metric coordinates using internal calibration**
- 4. Use triangulation equation to compute distance**
 - 1. If unknown baseline, simply invert disparity (reconstruction up to a scale factor)**
- 5. Post-process**
 - 1. remove outliers (e.g. median filter)**
 - 2. interpolate surface**

SOME OTHER MAJOR PROBLEMS WITH CORRESPONDENCE (2D VIEW)



OCCLUSION



**LIMITED FIELD
OF VIEW**

Other Problems:

- Photometric issues:
 - specularities
 - strongly non-Lambertian BRDF's
- Surface structure
 - lack of texture
 - repeating texture within horopter bracket
- Geometric ambiguities
 - as surfaces turn away, difficult to get accurate reconstruction (affine approximate can help)
 - at the occluding contour, likelihood of good match but incorrect reconstruction

MATCHING AND CORRESPONDENCE

There is no “Best” solution for correspondence

new frame-rate stereo systems use cross-correlation or SAD with left-right and right-left validation

other constraints include

- match ordering

- anomalous matches

- texture tests

- ambiguity (global minimum close to another local minimum)

There has been recent work on computing a “globally” optimal disparity map taking into account

- occlusion

- C^0 and C^1 discontinuities

- ordering constraints based on continuous surfaces

Results From B&C Paper

- Individual match definitions and results
- Heuristics used to prune matches
- Evaluation of combinations

Stereo Summary

- Geometry of simple non-verged system
- Epipolar geometry of two-camera system
- Use of rectification to reduce latter to former
- Match metrics
 - feature-based
 - region-based
- Match heuristics
- Some experimental results

GENERAL BINOCULAR STEREO SYSTEM

In general, the cameras will be in “general position” with more general internal structure.

Let's assume we know the full camera calibration (R, t, A)
First, we always convert from pixel to metric coordinates, then note that for one camera, we have

$$\begin{aligned}u(\mathbf{R}_z \mathbf{p} + \mathbf{t}_z) &= f(\mathbf{R}_x \mathbf{p} + \mathbf{t}_x) \\v(\mathbf{R}_z \mathbf{p} + \mathbf{t}_z) &= f(\mathbf{R}_y \mathbf{p} + \mathbf{t}_y) \quad \text{which we can re-organize as} \\ \mathbf{A} \mathbf{p} &= \mathbf{b}\end{aligned}$$

We get two such equations for the two stereo cameras leading to

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} p = A^* p = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = b^*$$

GENERAL BINOCULAR STEREO SYSTEM

We get two such equations for the two stereo cameras leading to

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} p = Mp = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = c$$

This is an overdetermined system with solution

$$\mathbf{p} = (\mathbf{M}^t \mathbf{M})^{-1} \mathbf{M}^t \mathbf{c}$$